Estimation du niveau sonore de sources d'intérêt au sein de mixtures sonores urbaines : application au trafic routier

Jean-Rémy Gloaguen

UNIVERSITE BRETAGNE SCIENCES LOIRE POUR L'INGENIEUR

CENTRALE NANTES

THESE DE DOCTORAT DE

L'ÉCOLE CENTRALE DE NANTES
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 602
Sciences pour l'Ingénieur
Spécialité : *Acoustique*

Par
Jean-Rémy Gloaguen

Estimation du niveau sonore de sources d'intérêt au sein de mixtures sonores urbaines : application au trafic routier

Soutenue à Nantes, le 03 octobre 2018
Unité de recherche : Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux, Unité Mixte de Recherche en Acoustique Environnementale

Rapporteurs :

Régis Marchiano	Professeur des Universités, Sorbonne Université
Emmanuel Vincent	Directeur de Recherche, Institut national de recherche en informatique et en automatique

Composition du Jury :

Présidente :	Catherine Lavandier	
Examinateurs :	Catherine Lavandier	Professeur des Universités, Université de Cergy-Pontoise
	Nicolas Misdariis	Chargé de Recherche, Institut de Recherche et Coordination Acoustique/Musique
Dir. de thèse :	Jean-François Petiot	Professeur des Universités, École Centrale de Nantes
Encadrant :	Arnaud Can	Chargé de Recherche, Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux
	Mathieu Lagrange	Chargé de Recherche, École Centrale de Nantes

Remerciements

Durant ces 3 années de thèse, de nombreuses personnes auront été à mes côtés pour m'aider, m'épauler et me soutenir dans mes travaux.

C'est donc tout naturellement que je remercie, en premier lieu, mon directeur de thèse, Jean-François Petiot, Professeur des Universités, qui aura apporté par son regard de précieux conseils et suggestions notamment pour la réalisation des tests statistiques du test perceptif.

Je remercie également, mes encadrants, Arnaud Can et Mathieu Lagrange, tous deux chargés de recherche, respectivement à l'UMRAE et au LS2N, pour leur précieuse aide au quotidien, leurs points de vues complémentaires et pour s'être montrés si disponibles tout au long de ces 3 années. Les nombreuses discussions menées avec vous sur mes travaux ou sur le milieu de la recherche m'auront beaucoup apporté et appris ce qu'est le métier de chercheur.

Je tiens à remercier également chaque membre du jury pour avoir accepté de participer à ma soutenance et d'évaluer mon travail.

Je suis reconnaissant à Jean-Julien Aucouturier, chercheur CNRS au sein du laboratoire Sciences et Technologies de la Musique et du Son, d'avoir accepté d'être membre de mon CSI et dont les discussions sur les tests statistiques m'ont beaucoup aidé, ainsi qu'à Cédric Févotte, Directeur de Recherche CNRS à l'IRIT, d'avoir pris le temps de discuter avec nous sur ce sujet et d'avoir partagé son expérience et son point de vue.

Un grand merci à Judicaël Picaut, directeur de l'UMRAE, et aux membres de cette unité, pour m'avoir accueilli dans leur équipe et m'avoir offert de si agréables conditions de travail durant les trois années de cette thèse. Une distinction particulière pour Vincent Gary, technicien, qui a participé aux enregistrements des passages de voitures sur la piste avec nous. Je n'oublie pas mon camarade de bureau, Pierre Aumond, ingénieur de recherche, dont les échanges sur la science et le monde auront toujours été intéressants.

J'ai une pensée à mes camarades de l'Ifsttar qui poursuivent ou finissent leur thèse, nos repas hebdomadaires entre doctorants m'auront toujours fait plaisir et apporté une bouffée d'air frais bien utile. Bonne continuation à eux !

Je salue mes anciens camarades de promotions de Master dont j'ai eu plaisir à les retrouver

au grès des différents congrès d'acoustique. J'espère que d'autres réunions viendront à l'avenir !

En vrac, je suis reconnaissant à ces artistes qui m'auront accompagné dans les oreilles pendant ces phases de rédaction et de codage sous le logiciel Matlab : Jean-Sébastien Bach, Kevin Morby, Paul Desmond, Kimio Eto, Franz Schubert, Baden Powell, Les Parvarim, Simeon ten Holt, Father John Misty, Marissa Nadler...

Je remercie ma famille, mes parents et ma soeur qui m'ont toujours soutenu durant mes études et m'ont permis d'en arriver là.

Pour finir, j'ai une grande pensée et tendresse à celle qui partage ma vie depuis plus de sept ans, à celle qui m'a soutenue et réconforté tout au long de cette thèse, à celle qui m'a, de nombreuses fois, écouté parler de détails techniques bien trop confus parfois pour être compréhensibles. À toi, Ingrid, je te remercie.

Table des matières

Remerciements **i**

Table des matières **iii**

Table des figures **vii**

Liste des tableaux **xiii**

Liste des publications **xvii**

Introduction **1**

1 Connaitre l'environnement sonore urbain : modèles de prédiction et mesures **7**
1.1 Définition formelle du problème 8
1.2 Utilisation de modèles prédictifs 10
1.2.1 Modèle d'émission du trafic routier 11
1.2.2 Modèle de propagation 12
1.2.3 Réaliser des cartes du bruit de trafic en ville 13
1.2.4 Vers la modélisation d'autres sources sonores ? 15
1.2.5 Limitations des modèles prédictifs 16
1.2.5.1 Limites liées à la modélisation de la source et de la propagation 16
1.2.5.2 Limites liées à la simulation et à la représentation 17
1.3 Utilisation de mesures acoustiques 19
1.3.1 Déploiement de réseaux de capteurs fixes 19
1.3.2 Mesures mobiles 21
1.3.3 Mesures participatives 22
1.3.4 Intérêts et limites des mesures faites en ville 24
1.4 Estimation du niveau sonore du trafic routier 26
1.5 Méthode proposée 27
1.6 Conclusion du chapitre 29

2 Méthodes de séparation des sources sonores **31**
2.1 Analyse Computationnelle de Scènes Auditives 32

2.2 Algorithme DUET . . . 33
2.3 Analyse en Composantes Indépendantes . . . 35
2.4 Factorisation en Matrices Non-négatives . . . 37
2.5 Détection d'évènements sonores . . . 38
2.6 Comparaison des approches . . . 39

3 La Factorisation en Matrices Non-négatives 41
3.1 Principe de fonctionnement de la Factorisation en Matrice Non-négatives . . . 41
3.2 Fonction de coût et familles de divergences . . . 44
3.3 Une sous-classe des divergences de Bregman : la β-divergence . . . 45
3.3.1 Distance Euclidienne . . . 46
3.3.2 Divergence de Kullback-Leibler . . . 47
3.3.3 Divergence d'Itakura-Saito . . . 47
3.3.4 Autres familles de divergences . . . 48
3.4 Mise à jour des formes de **W** et de **H** . . . 48
3.4.1 Algorithme heuristique par descente de gradient . . . 48
3.4.2 Algorithme multiplicatif par *majorisation-minimisation* . . . 49
3.4.2.1 Définition de la fonction auxiliaire . . . 50
3.4.2.2 Construction de la fonction auxiliaire . . . 51
3.4.3 Autres approches . . . 53
3.5 Analyse Probabiliste en Composantes Latentes . . . 53
3.6 Apprentissage du dictionnaire . . . 54
3.6.1 Apprentissage supervisé et non-supervisé . . . 55
3.6.2 Apprentissage semi-supervisé . . . 55
3.7 NMF initialisée seuillée . . . 57
3.8 NMF avec contraintes . . . 60
3.8.1 Contrainte de parcimonie . . . 60
3.8.2 Contrainte de régularité temporelle . . . 61
3.8.3 Autres contraintes . . . 64
3.9 Conlusion du chapitre . . . 65

4 Création de corpus de mixtures sonores urbaines 67
4.1 Création de scènes sonores : choix d'une méthode . . . 68
4.1.1 Auralisation d'ESU . . . 68
4.1.2 Simulateur de scènes sonores . . . 69
4.2 Présentation de *SimScene* . . . 71
4.3 Création d'un corpus élémentaire d'échantillons audio . . . 72
4.3.1 Recherche en ligne des échantillons audio . . . 72
4.3.2 Enregistrements de passages de véhicules . . . 73
4.3.3 Composition du corpus élémentaire complet . . . 74
4.4 Corpus d'évaluation *Ambiance* . . . 75

4.5 Corpus d'évaluation de scènes sonores urbaines réalistes . . . 78
4.5.1 Présentation des enregistrements audio de références . . . 78
4.5.2 Écoutes des scènes sonores . . . 79
4.5.3 Annotation des enregistrements sonores . . . 80
4.5.4 Reproduction des enregistrements audio . . . 83
4.6 Validation du réalisme du corpus d'évaluation *SOUR* par un test perceptif . . . 84
4.6.1 Mise en place du test . . . 85
4.6.2 Résultats . . . 88
4.6.2.1 Constitution du panel . . . 88
4.6.2.2 Distribution des notes des scènes enregistrées et répliquées . . . 89
4.7 Conclusion du chapitre . . . 93

5 Étude du comportement de la NMF sur le corpus d'évaluation *ambiance* 95
5.1 Rappel de la méthode employée . . . 96
5.2 Estimateur de référence . . . 97
5.3 Estimateur basé sur la NMF . . . 97
5.3.1 Constitution du dictionnaire . . . 98
5.3.2 Réalisation de la NMF . . . 99
5.3.3 Résumé des facteurs expérimentaux . . . 101
5.4 Performances de l'estimateur *baseline* . . . 102
5.5 Performances de l'estimateur basé sur la NMF . . . 105
5.5.1 Erreurs MAE_g . . . 106
5.5.2 Influence des facteurs expérimentaux w_t et K . . . 109
5.5.3 Influence de l'initialisation de la NMF IS . . . 109
5.5.4 Erreurs MAE_{TIR} et fonctions de coût . . . 110
5.5.5 Erreurs MAE pour chaque ambiance et valeur du TIR . . . 112
5.6 Conclusion du chapitre . . . 117

6 Performances de la NMF sur le corpus d'évaluation *SOUR* 121
6.1 Rappel de l'expérience menée . . . 122
6.2 Erreurs MAE_g obtenues par l'estimateur *baseline* . . . 125
6.3 Erreurs MAE_g obtenues par l'estimateur NMF . . . 126
6.4 Erreurs MAE_{60} par ambiance sonore . . . 127
6.5 Comparaison des niveaux sonores $L_{eq,tr.,1s}$ pour plusieurs scènes sonores. . . . 131
6.6 Pistes d'amélioration . . . 133
6.6.1 Contrainte de régularité temporelle . . . 133
6.6.2 Optimisation par les environnements sonores . . . 137
6.7 Conclusion du chapitre . . . 139

Conclusions générales et perspectives 141

Appendices **147**

A Développement de la contrainte de régularité temporelle pour $\beta = 0$ et $\beta = 2$ 149
A.1 Cas de la distance Euclidienne 149
A.2 Cas de la divergence d'Itakura-Saito 151

B Analyses complémentaires des résultats du test perceptif **157**
B.1 Effets des auditeurs sur l'évaluation des scènes 157
B.2 Effets de l'ambiance sonore 158

C Estimation du seuil optimal par des indicateurs sonores **161**

D Impact de la contrainte de parcimonie sur la corpus *SOUR* **167**

E Correspondance des noms de scènes du corpus *SOUR* **171**

Bibliographie **175**

Table des figures

1 Schéma de principe de la séparation de sources. 3

1.1 Schéma du problème considéré en ESU pour un signal capté par un microphone au point **x**. Deux sources sonores émises à l'instant $t - \tau_j$ à la position x_j sont présentes : une voiture, s_1 (résumée en une source ponctuelle symbolisée par un cercle rouge), et un piéton, s_2, (résumée en une source ponctuelle symbolisée par un cercle bleu). Chaque source se propage jusqu'au récepteur selon 2 chemins de propagation (champ direct δ_{ij1} et champ réfléchi δ_{ij2}). 9
1.2 Résumé des étapes menant aux cartes de bruit du trafic routier. 13
1.3 L_{DEN} (a) et L_N (b) de l'île de Nantes pour le trafic routier [Nantes Métropole, 2017]. 14
1.4 Schéma d'un réseau de capteurs fixes. 19
1.5 Captures d'écran de l'application *NoiseCapture*. 23
1.6 Carte de l'ESU de l'île de Nantes mesurée par l'application *NoiseCapture* (relevée le 22/03/2018). 24
1.7 Spectrogrammes d'un passage d'une voiture (a), d'un sifflement d'oiseaux (b), d'un klaxon (c) et d'un bruit de pas (d). 25
1.8 Vue synthétique de l'approche proposée. 26
1.9 Schéma de l'approche par séparation de sources d'un mélange sonore mélangeant une composante *trafic* avec des sons de klaxons. 27
1.10 Schéma-bloc du protocole expérimental. 28

2.1 Histogramme 2D pour un signal sonore composé de 5 sources sonores, générant 5 pics [Rickard, 2007]. 34

3.1 Exemple d'une NMF pour un signal audio de mixture urbaine composé de 3 sources sonores. **W** et **H** sont constitués de 3 bases (K = 3) : a) un spectre voiture, b) un spectre de klaxon, c) un spectre d'oiseau. 42
3.2 Évolution des β-divergences pour un cas simple ($x = 1$). 46
3.3 Schéma-bloc des étapes de la NMF. 54
3.4 Schéma-bloc des étapes de la NMF semi-supervisée. 56
3.5 Similarité cosinus pour une représentation linéaire et sigmoïdienne de la distance $D_\theta(\mathbf{W_0}\|\mathbf{W}')$ trié dans l'ordre décroissant avec un seuillage dur $t_h = 0{,}6$. 58

3.6 Pondération α appliquée à $\mathbf{W}'$, exprimé linéairement, composé de 100 bases avec pour seuil $t_h = 0,2$ (seuillage dur) et $t_{f,1} = 0,15$ et $t_{f,2} = 0,30$ (pour le seuillage *firm*). 59
3.7 Exemple de l'effet de la parcimonie pour une scène du corpus *alerte* pour $\alpha_{sp} \in \{0,\ 0,5,\ 5\}$. 62
3.8 Influence de la pondération α de la contrainte temporelle $C_t(\mathbf{H})$ sur la somme; selon K, des activateurs sur une scène audio de 30 secondes. 64

4.1 Représentation temporelle (à gauche), *Piano Roll* (au centre) et spectrogramme (à droite) générés par *SimScene* d'une scène composée, d'un bruit de fond *trafic* (en vert foncé) et *parc* (en gris) et d'évènements *oiseaux* (en vert), *voiture* (en magenta) et *passant* (en rouge). 72
4.2 Zoom du spectrogramme (nombre de points $w = 2^{12}$ avec 50 % de recouvrement) dans la bande de fréquence $[1500 - 7500]$ Hz d'un enregistrement de passage de véhicule (véhicule Renault, rapport 3, 40 km/h). À gauche, l'enregistrement original, à droite l'enregistrement filtré par le filtre médian. 74
4.3 Schéma bloc de la pondération du signal trafic selon la scène i et le TIR. 76
4.4 Spectres sonores moyens des classes *interférante* (courbes en bleu) et *trafic* (courbes en rouge) pour chaque sous-corpus. 77
4.5 Exemple de l'évolution du niveau sonore équivalent 1 seconde $L_{eq,1s}$ d'une mixture sonore extraite du sous-corpus *alerte* avec la composante *trafic* calibrée à $TIR \in \{-12, 0, 12\}$ dB. 78
4.6 Schéma bloc résumant la création du corpus d'évaluation de scènes sonores urbaines réalistes *SOUR*. 79
4.7 Parcours réalisé avec les 19 points de mesures avec le niveau sonore mesuré équivalent. 79
4.8 Valeurs du TIR par scène et moyennés par ambiance sonore pour le corpus *SOUR*. 84
4.9 Distribution des scènes audio pour chaque juge selon leur type (enregistré ou répliqué) : en bleu la quantité de scènes enregistrées évaluées par le juge et en rouge le nombre de scènes répliquées. La somme de ces deux parties équivaut au nombre de scènes testées K par juge. 87
4.10 Nombre de réplications, R, pour chaque scène obtenu dans X_{opt} avec comme combinaison $J = 50$, $B = 40$, $K = 20$. Les 20 premières scènes sont les scènes issues des enregistrements du projet GRAFIC, les 20 suivantes sont les scènes répliquées sous *SimScene*. 88
4.11 Résumé des informations relatifs aux auditeurs. 89
4.12 Représentation en diagramme en boîte à moustache entre les scènes enregistrées et répliquées. 90
4.13 Boites à moustaches pour les scènes enregistrées (a) et pour les scènes répliquées (b) classées selon leur ambiance sonore. 92

5.1 Schéma-bloc de l'estimation du niveau sonore du bruit de trafic. 96
5.2 Principe de l'estimateur *baseline* pour une mixture sonore filtrée à $f_c = 5$kHz. . . 97
5.3 Schéma bloc de l'estimateur NMF sur le corpus d'évaluation *Ambiance*. 98
5.4 Schéma bloc de la création du dictionnaire. 98
5.5 Création des éléments de $\mathbf{W}$ sur un extrait de 3 secondes du passage d'une voiture pour une trame temporelle $w_t = 1$ seconde. À gauche, le spectrogramme du signal audio avec, en pointillés, une fenêtre de découpe. Au centre, le signal découpé en trois trames et dont les valeurs *rms* sont ensuite calculées générant 3 spectres. . . 99
5.6 Spectre en fréquence du passage d'une voiture en bandes fines (2049 points) et en bandes de tiers d'octave (29 bandes). 100
5.7 Organigramme des différents facteurs expérimentaux impliqués dans l'outil *estimateur*. 102
5.8 Distributions des erreurs relatives entre les niveaux estimés et exactes pour chaque sous-corpus et chaque TIR pour l'estimateur filtre passe-bas à la fréquence de coupure $f_c = 500$ Hz. 103
5.9 Erreurs MAE pour chaque sous-corpus et chaque TIR pour l'estimateur filtre passe-bas à la fréquence de coupure $f_c = 500$ Hz. 104
5.10 Distributions des erreurs relatives entre les niveaux estimés et exactes pour chaque sous-corpus et chaque TIR pour l'estimateur filtre passe-bas à la fréquence de coupure $f_c = 20$ kHz. 105
5.11 Erreurs MAE pour chaque sous-corpus et chaque TIR pour l'estimateur filtre passe-bas à la fréquence de coupure $f_c = 20$ kHz. 105
5.12 Distances moyennes $D_\theta(\mathbf{W_0}\|\mathbf{W}')$ triées par ordre décroissant pour chaque TIR, obtenues pour $w_t = 0{,}5$ s, $K = 200$ et $\beta = 1$ et $t_h = 0{,}41$. 108
5.13 Influence de la forme du dictionnaire pour les 3 versions optimales de la NMF retenues. 109
5.14 Distances moyennes $D_\theta(\mathbf{W_0}\|\mathbf{W}')$ triées par ordre décroissant pour chaque TIR, obtenues pour $w_t = 0{,}5$ s, $K = 200$ et $\beta = 1$ pour un dictionnaire initialisé par des valeurs aléatoires. 111
5.15 Evolution de la fonctions de coût $D(\mathbf{V}\|\mathbf{WH})$ et de l'erreur MAE_g moyennes pour les combinaisons optimales des NMF SUP, SEM et IS sur l'intégralité du corpus d'évaluation *Ambiance*. 111
5.16 Distributions des erreurs relatives pour chaque classe pour la NMF SUP (a), SEM (b) et IS (c). 113
5.17 Erreurs MAE pour chaque classe pour la NMF SUP (a), SEM (b) et IS (c). . . . 114
5.18 Comparaisons des niveaux sonores équivalents des classes *trafic* et *interférante* pour 2 valeurs du TIR (-12 dB et 12 dB) pour la scène 25 du sous-corpus *alerte*. 115
5.19 Comparaisons des spectres des classes *trafic* et *interférante* avec la somme des 2 éléments de $\mathbf{W_r}$ pour 2 valeurs du TIR (-12 dB et 12 dB) pour 2 sous-corpus (*alerte* et *climat)*. 116

5.20 Évolution de l'erreur MAE pour chaque sous-corpus selon la valeur seuil t_h pour $TIR = -12$ dB (5.20(a)), $TIR = 0$ dB (5.20(b)) et $TIR = 12$ dB (5.20(c)). . . . 118

6.1 Schéma-bloc des étapes dans l'estimation du niveau sonore du trafic pour le corpus d'évaluation *SOUR*. . . . 122
6.2 Estimation des niveaux sonores équivalents pour une durée d'intégration de 60 secondes sur deux scènes. Le deuxième niveau sonore $L_{eq,tr.,60s,2}$ inclut 9 secondes de la scène 1 et 51 secondes de la scène 2. . . . 124
6.3 Comparaison de l'évolution de la fonction de coût et des erreurs MAE_g moyennes selon les 3 NMF optimales retenues. . . . 127
6.4 Distribution des erreurs relatives selon les 3 NMF optimales retenues et les 4 ambiances sonores. . . . 128
6.5 Distances $D_\theta(\mathbf{W_0}\|\mathbf{W}')$ moyennes par ambiance sonore triées par ordre décroissant dans le cas de la NMF IS optimale ($\beta = 2$, $K = 300$, $w_t = 1$). . . . 129
6.6 Évolution de l'erreur MAE_g et du temps de calcul selon la taille des matrices K dans le cas de la NMF IS optimale (pour la figure, le nombre d'éléments dans K a été étendu à 400). . . . 130
6.7 Évolution du niveau sonore équivalent $L_{eq,1s}$ durant 60 secondes de la scène *parc-05*. 131
6.8 Évolution du niveau sonore équivalent $L_{eq,1s}$ durant 60 secondes de la scène *Rue calme-07*. . . . 132
6.9 Évolution du niveau sonore équivalent $L_{eq,1s}$ durant 60 secondes de la scène *Rue bruyante-03*. . . . 133
6.10 Évolution du niveau sonore équivalent $L_{eq,1s}$ durant 60 secondes de la scène *Rue très bruyante-05*. . . . 133
6.11 Influence de la pondération de régularité temporelle pour la NMF SEM selon chaque ambiance sonore. . . . 135
6.12 Comparaisons des spectres de la partie libre $\mathbf{W_r H_r}$ pour le cas sans pondération ($\alpha_t = 0$) et avec ($\alpha_t = 0,05$) pour la scène 2 de l'ambiance *Parc* (a) et la scène 7 de *Rue très bruyante* (b). . . . 136
6.13 Impact de la contrainte de régularité temporelle sur la fonction de coût et les erreurs MAE_g moyennes pour la NMF SEM avec $\beta = 0$, $w_t = all$ s et $K = 50$. . . 137
6.14 Influence de la contrainte de régularité temporelle sur la NMF IS pour leur combinaison optimale de modalités ($\beta = 1$, $w_t = 1$, $K = 200$, $t_h = 0{,}36$) pour $\alpha_t = 0$ et $\alpha_t = 0{,}5$. . . . 137
6.15 Influence de la valeur seuil t_h sur les erreurs MAE_{60} selon l'ambiance sonore. . . 138

A.1 Influence de la contrainte de régularité temporelle pour $\beta = 2$ selon l'algorithme obtenu par *majorisation-minimisation*. . . . 151
A.2 Influence de la contrainte de régularité temporelle pour $\beta = 0$ selon l'algorithme obtenu par *majorisation-minimisation*. . . . 155

B.1 Diagramme des effets d'intéraction type∗ambiance. 159

C.1 Évolution des seuils optimaux et des indicateurs Δ_L selon les ambiances. 163
C.2 Corrélations entre l'évolution des seuils optimaux et des indicateurs $\Delta_{L_x-L_y}$. . . 164
C.3 Schéma de principe de la détermination du niveau sonore du trafic par seuil interpolé. 164
C.4 Exemple d'interpolation réalisée pour une scène avec un indicateur $\Delta_{L_{1k}-L_{5k}} = 11 : t_{h,interp} = 0{,}343$. 165
C.5 Influence de la méthode de seuillage (seul fixe t_h, optimisé $t_{h,opt.}$, interpolé $t_{h,interp.}$) à partir de l'indicateur $\Delta_{L_{1k}-L_{5k}}$ sur les erreurs MAE_{60}. 166

D.1 Influence de la parcimonie sur l'erreur MAE_{60} selon la NMF SEM optimale avec et sans pondération. 169
D.2 Comparaison des spectres de la classe *interférante* avec la somme des 2 éléments de $\mathbf{W_r}$ pour le cas sans pondération ($\alpha_{sp} = 0$) et avec ($\alpha_{sp} = 0,05$) pour la scène 2 de l'ambiance *Parc* (a) et la scène 7 de *Rue très bruyante* (b). 169

Liste des tableaux

1.1 Paramètres d'estimation de la puissance acoustique selon 3 modèles d'émission sonore. 12

2.1 Comparaison des 4 méthodes de séparation de sources et de la méthode DESA en vue d'estimer le niveau sonore du trafic routier à partir d'enregistrements monophoniques (« - » : pas adapté, « + » : adapté, « ++ » : très adapté). 39

3.1 Fonctions concaves, convexes et constantes selon β. 51

4.1 Ensemble de mesures réalisées sur pistes avec des passages de véhicules à vitesses stabilisée (à gauche) et en accélération et freinage (à droite). 73

4.2 Composition de la base de données pour les évènements sonores. 75

4.3 Composition de la base de données pour les bruits de fond. 75

4.4 Résumé des classes de sons inclues dans les classes interférantes, seules les classes *alerte* et *transport* ne contiennent pas de bruit de fond. 76

4.5 Résumé des 19 points de mesures avec l'ambiance générale [Aumond *et al.*, 2017a]. 80

4.6 Classification des scènes par ambiances sonores. 80

4.7 Exemple d'un fichier d'annotation pour la scène 1-EW-07. 81

4.8 Niveau sonore et description des classes de sons les plus récurrentes dans l'environnement urbain (nombre d'évènements sonore par minute $> 0{,}1$/min). 82

4.9 Durées cumulées par ambiance du corpus *SOUR*. 83

4.10 Résumé des 40 audio composant l'ensemble des scènes testées avec les temps d'extraction des 30 secondes d'audio, l'identifiant et le nom des fichiers audio originaux. 85

4.11 DDL, valeurs t et valeurs p pour chaque test de Student mené entre les scènes enregistrées et répliquées; en gras, les valeur p supérieures au seuil de signification de 5 %. 91

5.1 Facteurs expérimentaux et leurs modalités utilisés pour le corpus d'évaluation *Ambiance*. 101

5.2 Erreurs MAE_g et MAE_{TIR} en dB de l'estimateur *baseline* selon f_c sur l'ensemble du corpus *Ambiance* et pour chaque TIR. En gras-rouge l'erreur MAE_g la plus faible, en gras-noir, les erreurs MAE_{TIR} les plus faibles selon les fréquences f_c. . 103
5.3 Erreurs MAE_g les plus faibles de la NMF SUP et NMF SEM pour le corpus d'évaluation *Ambiance*, en gras-rouge, l'erreur globale la plus faible. 106
5.4 Erreurs MAE_g les plus faibles de la NMF IS pour le corpus d'évaluation *Ambiance* selon la représentation linéaire ou sigmoïde de la distance $D_\theta(\mathbf{W_0}\|\mathbf{W}')$. 107
5.5 Erreurs MAE_g les plus faibles de la NMF IS pour le corpus d'évaluation *Ambiance* selon un seuillage dur ou *firm*. 108
5.6 Erreurs MAE_g les plus faibles de la NMF IS pour le corpus d'évaluation *Ambiance* selon l'initialisation du dictionnaire. En ligne 1 et 3, le seuil est celui permettant d'obtenir l'erreur la plus faible, en ligne 2, le seuil correspond à celui obtenu avec une initialisation par $\mathbf{W_0}$ mais sur une NMF IS initialisée par des valeurs aléatoires. 110
5.7 Erreurs MAE_{TIR} en dB selon les combinaisons optimales de la NMF SUP, SEM et IS avec l'estimateur *baseline* $f_c = 500$ Hz et les erreurs pour le filtre passe-bas $f_c = 20$ kHz. 112

6.1 Facteurs expérimentaux et leurs modalités utilisés pour le corpus *SOUR*. 123
6.2 Corpus d'évaluation *SOUR* par ambiance sonore selon le nombre de scènes N, leur durée totale et le nombre de niveaux sonores calculées N_{60}. 125
6.3 Erreurs moyennes MAE_g et MAE_{60} en dB pour l'estimateur *baseline* pour le corpus d'évaluation *SOUR*, en gras-rouge la plus faible erreur MAE_g, en gras-noir les erreurs MAE_{60} les plus faibles. 125
6.4 Erreurs MAE_{60} en dB les plus faibles pour les combinaisons optimales de modalités des estimateurs pour le corpus d'évaluation *SOUR*. 126
6.5 Erreurs MAE_{60} en dB les plus faibles selon les estimateurs NMF pour chaque méthode dans sa combinaison optimale de modalités avec les estimateurs *baseline* à 500 Hz et 20 kHz. En gras-rouge, les erreurs globales les plus faibles, en gras-noir, les erreurs de la NMF inférieures à l'estimateur *baseline* $f_c = 500$ Hz. 128
6.6 Erreurs MAE_{60} les plus faibles pour les combinaisons optimales de modalités des estimateurs pour le corpus d'évaluation *SOUR* en présence d'une pondération de régularité temporelle. 135
6.7 Erreurs MAE_{60} minimales selon le seuil optimal $t_{h,opt}$ par ambiance sonore, en gras les erreurs minimales. 138

B.1 Résultats de l'ANOVA avec interaction avec les facteurs *type* et *auditeur*. 158
B.2 Résultats de l'ANOVA avec interaction avec les facteurs *type* et *ambiance*. 158

C.1 Erreurs MAE_{60} minimales selon le seuil fixe t_h et optimal $t_{h,opt}$ par ambiance sonore. 161
C.2 Influence de l'indicateur d'optimisation dans l'estimation de l'erreur MAE_g. . . . 166

D.1 Erreurs MAE_{60} les plus faibles pour les combinaisons optimales des modalités des estimateurs pour le corpus d'évaluation *SOUR* en présence d'une pondération de parcimonie. 168

E.1 Correspondances des noms des scènes enregistrées et répliquées pour l'ambiance *Parc*. 171
E.2 Correspondances des noms des scènes enregistrées et répliquées pour l'ambiance *Rue calme*. 172
E.3 Correspondances des noms des scènes enregistrées et répliquées pour l'ambiance *Rue bruyante*. 173
E.4 Correspondances des noms des scènes enregistrées et répliquées pour l'ambiance *Rue très bruyante*. 173

Liste des publications

Publications dans des revues d'audience internationale à comité de lecture

- Gloaguen, J.-R., Lagrange, M., Can, A., et Petiot, J. F. (2018c). Estimation of the road traffic sound levels in urban areas based on non-negative matrix factorization techniques. *Journal on Audio, Speech and Music Processing*. en révision Journal on Audio, Speech and Music Processing. en révision
- Gloaguen, J.-R., Can, A., Lagrange, M., et Petiot, J. F. (2018b). Road traffic sound level estimation from realistic urban sound mixtures by non-negative matrix factorization. *Applied Acoustics*, 143(1), 229–238.

Communications à des congrès internationaux à comité de sélection et actes publiés

- Gloaguen, J. R., Can, A., Lagrange, M., et Petiot, J. F. (2017, June). Creation of a corpus of realistic urban sound scenes with controlled acoustic properties. In *Proceedings of Meetings on Acoustics 173EAA* (Vol. 30, No. 1, p. 055009). ASA.
- Gloaguen, J.-R., Can, A., Lagrange, M., et Petiot, J.-F. (2016a). Estimating traffic noise levels using acoustic monitoring : A preliminary study. In *DCASE 2016, Detection and Classification of Acoustic Scenes and Events*, 4p

Communications à des congrès nationaux et actes publiés

- Gloaguen, J.-R., Can, A., Lagrange, M., et Petiot, J. F. (2018a). Estimation du niveau sonore du trafic routier au sein de mixtures sonores urbaines par la factorisation en matrices non négatives. In *14ème Congrès Français d'Acoustique CFA '18 Le Havre*, pp 286-291

Autres

- Gloaguen, J.-R., Can, A., Lagrange, M., et Petiot, J. F. (2017b). Reconnaissance et estimation du bruit de trafic dans l'environnement urbain. *Acoustique et Techniques : trimestriel*

d'information des professionnels de l'acoustique, 86 :54–60

- Gloaguen, J.-R., Can, A., Lagrange, M., and Petiot, J. F. (2016b). Estimation du niveau sonore du trafic routier par mesures acoustiques : résultats préliminaires. In *Journée Jeunes Chercheurs en Acoustique musical, Audio et Signal*

Introduction

Au sein de l'Union Européenne, 70 % de la population, soit quasiment 340 millions d'habitants, vivent dans des zones urbaines [WHO, 2017]. 486 villes concentrent, chacune, plus de 100 000 habitants. En France, selon l'INSEE, c'est même plus de 84 % de la population qui vit dans une zone urbaine, soit plus de 55 millions d'habitants. Cette concentration soulève de grandes questions autour de l'organisation de l'espace urbain afin d'offrir une qualité de vie acceptable aux citadins. En effet, avec de telles densités (environ 3000 habitants/km^2 et jusqu'à plus de 21 000 habitants/km^2 pour la ville de Paris, la plus dense de l'Union Européenne (UE)), plusieurs formes de pollutions viennent dégrader l'environnement urbain. Des sources de désagrément perçues par le citadin, le bruit est le phénomène qui provoque le plus de gêne après la pollution de l'air. Ce bruit est le fruit des activités humaines, provenant essentiellement du transport qu'il soit routier, ferroviaire ou aérien [Zannin *et al.*, 2013]. En France, selon un rapport de l'Agence de l'Environnement et de la Maîtrise de l'Énergie [EUROPEENS, 2016], ce sont 52 millions de personnes qui se disent affectées par le bruit et principalement le bruit issu du trafic routier. Plus de 7 millions d'individus sont exposés à des niveaux supérieurs à 65 dB(A) au quotidien et à plus de 55 dB(A) la nuit. Cette exposition quotidienne, à de tels niveaux, n'est pas sans conséquence pour l'être humain. L'impact de l'exposition du bruit sur l'organisme est observé et étudié depuis de nombreuses années [Ising *et al.*, 1980]. Parmi les effets possibles, les troubles du sommeil [Pirrera *et al.*, 2010], de la vigilance et de la concentration, l'augmentation du stress, de la pression artérielle et du rythme cardiaque [Babisch *et al.*, 2005, Babisch, 2008] sont les plus relevés. Cet impact sur la santé a également un coût financier pour la société : en France, ce coût est estimé à plus de 11,5 milliards d'euros par an dont une grande partie (89 %) est imputable au bruit du trafic routier [EUROPEENS, 2016]. De plus, si le bruit en ville impacte la vie des citadins, celui-ci se fait également ressentir auprès de la faune sauvage [Dutilleux, 2012, Francis *et al.*, 2009] leur causant également du stress ou en compliquant la communication entre les individus et leur reproduction.

À partir de ce constat, il est donc nécessaire et utile de savoir caractériser les environnements sonores urbains (ESU) afin d'estimer les sources sonores présentes, leurs niveaux sonores et leurs répartitions pour ainsi réduire et limiter leurs impacts sur les populations urbaines. Actuellement les outils les plus répandus pour étudier les ESU se basent sur des modèles prédictifs d'émission sonore et de propagation de 4 sources de bruit : le trafic routier, ferroviaire,

aérien et les Installations Classées pour la Protection de l'Environnement [1]. Ces modèles sont notamment utilisés dans le cas de la cartographie du bruit de trafic imposée par la directive Européenne 2002/49/CE [Parlement Européen, 2002]. Ces cartes de bruit permettent l'estimation des niveaux sonores équivalents pondérés A à travers l'ensemble de grandes villes afin de déterminer les lieux où les niveaux sonores sont élevés (et où des travaux d'aménagement sont nécessaires) ou bien ceux préservés par ces sources de bruits. Toutefois, l'utilisation de ces modèles présente plusieurs limites. Déjà, ces modèles se restreignent à 4 sources de bruit là où l'ESU est un mélange de nombreuses autres sources sonores. Également, l'évolution des niveaux sonores dans le temps n'est pas calculé alors que ceux-ci évoluent constamment. Aussi, ils ne permettent pas d'appréhender l'ESU à travers la perception qu'en ont les citadins [Aumond *et al.*, 2017a]. Mais surtout, ces modèles prédictifs offrent des estimations des niveaux sonores approximées basé sur une simplifications des émissions sonores des véhicules et des environnements urbains, qui ne peuvent pas être validées par des mesures. Afin d'avoir une connaissance plus fine et précise des ESU, plusieurs projets s'intéressent actuellement au déploiement de capteurs acoustiques en ville [Picaut *et al.*, 2017, Zambon *et al.*, 2017]. Ces réseaux s'insèrent dans le cadre de l'émergence de la ville intelligente (*Smart city*) et de l'Internet des objets (*Internet of things*) où il devient possible de générer de nombreuses mesures directement en ville à partir de réseaux de capteurs par exemple. Cette approche permet de considérer l'ensemble des sources sonores présentes en ville et l'effet de l'architecture de la ville sur la propagation du son. Les applications sont alors diverses : amélioration des cartes de bruits par assimilation des niveaux sonores calculés et mesurés, cartographie des environnements sonores, prise en compte de la perception du citadin, détection d'évènements sonores particuliers... La réalisation de mesures nécessite toutefois des outils de traitement du signal adaptés afin de pouvoir caractériser les sources sonores présentes, notamment par l'estimation de leur niveau sonore. Sans cette étape, la prise en compte de l'ensemble des sources sonores, sans distinction, est susceptible de mener à des erreurs d'interprétation. L'étude des contributions des différentes sources sonores, à partir d'enregistrements sonores réalisés dans un environnement urbain, a pour l'instant été peu étudiée. La ville est composée de différentes ambiances sonores (parc, quartier résidentiel, boulevard...) ainsi que de nombreuses sources sonores variées (voiture, oiseaux, bruit de pas et de voix, klaxons, fontaine...), susceptibles d'émettre du bruit simultanément. La formation d'un outil de caractérisation des sources adapté à cette diversité n'est alors pas trivial. C'est pourquoi il est nécessaire de restreindre, dans un premier temps, l'étude à une source sonore. Puisque le trafic routier est la source de bruit la plus gênante et que les applications liées à sa cartographie par la mesure sont celles qui, à l'heure actuelle, présentent le plus d'intérêt [Jagniatinskis et Fiks, 2014], celle-ci sera la source sonore d'intérêt de ces travaux.

En conséquence, l'objectif de cette thèse est de créer un outil permettant de déterminer

1. Les Installations Classées pour la Protection de l'Environnement résument les équipements qui peuvent générer des dangers ou des inconvénients auprès des populations humaines, de la nature, de l'agriculture...On y inclut, par exemple, les stations-services, les éoliennes de plus de 12 m de haut, les exploitation agricole, les installation de stockage de déchets...

la contribution sonore du trafic routier en ville. Cette composante sonore correspond au bruit de fond routier généré par le flot continu de véhicules et par le passage d'un véhicule émergeant. Dans le cadre d'enregistrements sonores réalisés par des capteurs individuels, la tâche est complexe car il faut réussir à déterminer la composante du signal *trafic* à partir d'un seul microphone. Si l'étude de sons environnementaux est de plus en plus traitée via, par exemple, les challenges DCASE [Stowell *et al.*, 2015, Mesaros *et al.*, 2017] avec la détection de signaux liés au transport ou à la classification de scènes sonores, l'étude et la détermination du niveau sonore de la composante *trafic* parmi des mixtures sonores restent peu étudiées. Récemment, [Leiba *et al.*, 2017] réussit à extraire le signal audio du trafic parmi des enregistrements sonores et à suivre la trajectoire du véhicule à partir d'une antenne acoustique composée de 128 microphones et des méthodes d'apprentissages. Dans le cas de capteurs monophoniques, une approche choisie dans [Socoró *et al.*, 2017] consiste à détecter la présence des autres sources sonores pour chaque trame temporelle. Lorsque la source détectée n'est pas reconnue comme *trafic*, celle-ci est rejetée et non considérée dans l'estimation du niveau sonore du trafic routier. L'approche que nous considérons dans ce travail est différente puisqu'elle implique une méthode de séparation de sources, dont le principe est présenté en Figure 1. Cette méthode consiste à isoler la contribution d'une seule source parmi une mixture sonore composée de plusieurs sources sonores. Plusieurs applications dans l'audio ont été développées dans le domaine de la musique [Smaragdis et Brown, 2003, Virtanen, 2007] ou pour de la voix [Weninger *et al.*, 2012, Yilmaz et Rickard, 2004]. Une fois la source sonore isolée il devient possible d'obtenir de nombreuses informations, et notamment le niveau sonore. Parmi les différentes méthodes existantes, il est nécessaire que celle choisie soit adaptée aux réseaux de capteurs monophoniques et qu'elle prenne facilement en compte le recouvrement temporel des sources sonores.

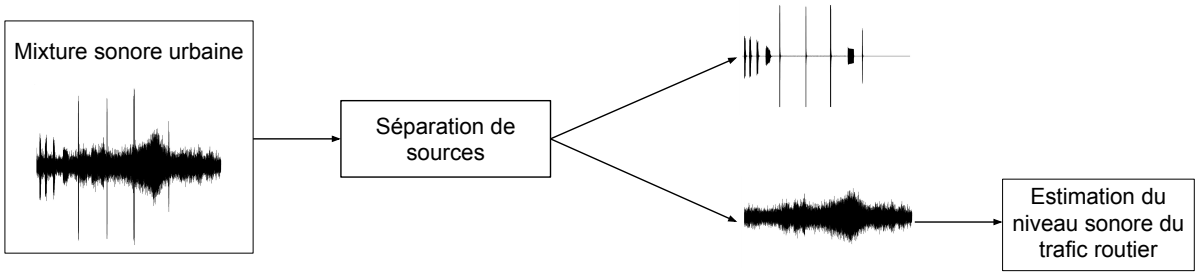


FIGURE 1 – Schéma de principe de la séparation de sources.

En cela, la Factorisation de Matrices Non-négatives (abrégée NMF pour *Non-negative Matrix Factorization* en anglais) [Lee et Seung, 1999] est l'approche retenue dans ces travaux car elle répond bien aux contraintes posées. Là encore, la NMF a été utilisée de très nombreuses fois pour des signaux contenant de la musique [Helen et Virtanen, 2005, Févotte *et al.*, 2009] ou de la parole [Wilson *et al.*, 2008a, Schmidt et Olsson, 2006]. Plusieurs variantes à cette méthode ont été développées dont certaines sont implémentées dans ces travaux. Son application sur une telle source n'ayant jamais encore été réalisée, son fonctionnement, face à de tels environnements

sonores, nécessite d'être étudié. Également, dans le cadre de cette thèse, une nouvelle forme de NMF est proposée, appelée NMF *initialisée seuillée*. Afin d'étudier ses performances, la NMF n'est pas appliquée sur des enregistrements audio, car l'estimation faite du niveau sonore du trafic ne pourrait alors pas être comparée à une valeur exacte, mais sur des corpus de scènes sonores simulées. L'intérêt de ce procédé est qu'il permet de connaitre les contributions de chacune des sources dont la composante *trafic*. La comparaison du niveau exact et estimé est alors possible.

Le propos de cette thèse est donc de proposer un protocole expérimental permettant d'évaluer la qualité d'estimation du niveau sonore du trafic routier par séparation de sources sur des mixtures simulées de scènes sonores urbaines. En conséquence, plusieurs approches de la NMF sont étudiées afin de définir l'approche optimale permettant d'obtenir les erreurs d'estimation les plus faibles.

Pour cela, il est nécessaire de définir formellement, dans un premier chapitre, l'objet d'étude de ces travaux, l'environnement sonores urbain, et de présenter les différentes méthodes visant à le caractériser, avec un exercice critique de leurs avantages et de leur limites. De ces observations faites, une proposition décrivant le protocole expérimental mis en place est alors faite.

Le second chapitre a pour objectif de décrire les différentes méthodes de séparation de sources qui peuvent être envisagées. Ces méthodes sont ensuite comparées selon le cahier des charges défini. La Factorisation en Matrices Non-négatives est alors la méthode retenue. Son fonctionnement est présenté en détail dans le chapitre 3 selon les différentes méthodes d'apprentissage ou expressions des fonctions de coût. Ce chapitre introduit également la NMF *initialisée seuillée*, élaborée durant les travaux de ce doctorat.

Le chapitre 4 est dédié à la réalisation de deux corpus d'évaluation composés de scènes sonores simulées et à la formation d'une base de données de sons isolés permettant de les composer. Un premier corpus, nommé *Ambiance*, est construit en mélangeant artificiellement une composante *trafic*, dont le niveau sonore est calibré, avec d'autres classes de sons spécifiques. Ce premier corpus a pour vocation de tester la NMF, d'étudier son fonctionnement et ses performances face à des sons urbains. Le second corpus, nommé *SOUR* pour Scènes sOnores Urbaines Réalistes, se base sur des enregistrements audio qui ont été réalisés en ville. L'annotation de ces enregistrements permet de définir une partition temporelle qui sert à la construction des scènes sonores. Le rendu obtenu est alors soumis à un test perceptif visant à évaluer le réalisme sonore des scènes simulées.

Les chapitres 5 et 6 sont ensuite dédiés à l'étude des performances de la NMF soumise aux deux corpus d'évaluation. Dans un premier temps, le fonctionnement de la NMF face au corpus *Ambiance* est présenté en détail afin d'étudier le comportement de cette méthode face à de telles mixtures sonores. Les erreurs produites sur l'intégralité du corpus, selon le niveau sonore du trafic et selon chaque classe de son, sont présentées. Les résultats sur le corpus de *SOUR* sont ensuite exposés afin d'obtenir la forme optimale de NMF qui pourrait être implémentée dans des capteurs embarqués et utilisée comme outil de traitement du signal *trafic*. Des méthodes d'optimisation sont alors proposées afin d'améliorer les performances de la NMF pour la tâche

visée.

Chapitre 1

Connaitre l'environnement sonore urbain : modèles de prédiction et mesures

Résumé

La description des environnements sonores urbains (ESU) est, à l'heure actuelle, réalisée par plusieurs outils dont les plus répandus sont basés sur des modèles d'émissions sonores et de propagation, notamment utilisés pour la cartographie de bruit de trafic. Ces outils possèdent plusieurs limites comme le nombre de sources considérées ou les approximations générés par ces modèles. Afin d'obtenir une meilleure description des environnements sonores, l'utilisation de mesures et d'enregistrements sonores en ville, à travers des réseaux de capteurs fixes et des mesures mobiles, est de plus en plus envisagée. Cette approche permet de considérer l'ESU dans son intégralité en prenant en compte l'ensemble des sources et les effets de l'environnement urbain sur leur propagation. Toutefois, pour utiliser pleinement ces mesures, il reste à savoir estimer les contributions des différentes sources présentes notamment le trafic routier. La méthode et le protocole expérimental mis en place pour ces travaux sont alors exposés.

Dans ce chapitre, une présentation des méthodes utilisées pour caractériser l'environnement sonore urbain est réalisée. Dans une première partie, le problème général est posé formellement, puis l'utilisation de modèles prédictifs et les éléments de cartographie de bruit en ville sont exposés et enfin la réalisation de mesures en milieu urbain est présentée. Enfin, la problématique générale et la solution proposée sont exposées.

1.1 Définition formelle du problème

Soit $M_i(t)$, un environnement sonore urbain (abrégé ESU) défini dans un espace Ω, capté en un point donné $i_{\in\Omega}$, reçu à un instant t. L'ESU se décompose alors comme la somme de N différentes contributions sonores $S_j(t)$ reçues à ce point i. Chacune de ces contributions est le résultat de l'émission sonore d'une source acoustique s_j, de puissance sonore $L_{w,j}$, située à la position $j_{\in\Omega}$ et émise à un instant $t - \tau_j$, qui s'est propagée dans l'environnement urbain le long de k chemins de propagation. L'ESU s'exprime alors, mathématiquement, dans le domaine temporel comme :

$$M_i(t) = \sum_{j=1}^{N} S_j(t), \tag{1.1a}$$

$$= \sum_{j=1}^{N} s_j(t - \tau_j) * \delta_{ij}(t), \tag{1.1b}$$

$$= \sum_{j=1}^{N} \sum_{k=1}^{+\infty} s_j(t - \tau_j)\delta_{ijk}. \tag{1.1c}$$

Dans l'équation 1.1b, le produit de convolution de la source s_j par la variable $\delta_{ij}(t)$ qui correspond à la réponse impulsionnelle de l'environnement urbain entre le point $M_i(t)$ et la source s située au point j, traduit l'intégralité des effets de propagation générés par la diffusion de s_j dans l'environnement Ω. Ils incluent les phénomènes d'atténuation géométrique de l'onde sonore ainsi que ceux de diffusion et d'absorption provoqués par les réflexions de l'onde sonore sur les parois des bâtiments et sur le sol. En conséquence, l'équation 1.1c décompose, pour chaque source j, l'impact de chaque chemin k de propagation de l'onde sonore. Ces chemins incluent le champ direct, le champ réfléchi par une réflexion, deux réflexions... Pour chaque champ, comme la distance de propagation est différente, le temps de propagation entre la source et le récepteur varie créant une atténuation δ_{ijk} spécifique. Le mélange $M_i(t)$ peut alors soit être captée par un microphone installé en ville (comme dans l'exemple en Figure 1.1), ce qui permet notamment d'obtenir des indicateurs physiques (niveau sonore en dB SPL ou pondéré A), soit être perçue par les citadins où les aspects perceptifs entrent alors en jeux. Dans ce cas, le mélange $M_i(t)$ est évaluée à travers des indicateurs perceptifs comme l'*agrément sonore* qui dépend notamment de la prédominance de certaines sources sonores et du cadre environnemental dans lequel elles sont perçues. L'ensemble de ces variables est représenté dans la Figure 1.1.

Les sources s_j expriment les différentes sources sonores présentes en ville comme le trafic aérien ou ferroviaire mais aussi les voix, les bruits de pas, celui d'une valise à roulettes, les sifflements d'oiseaux, les aboiements de chiens... La source sonore principale en ville est celle du trafic routier, $S_{tr.}(t)$, qui est considérée comme la somme des contributions des M véhicules

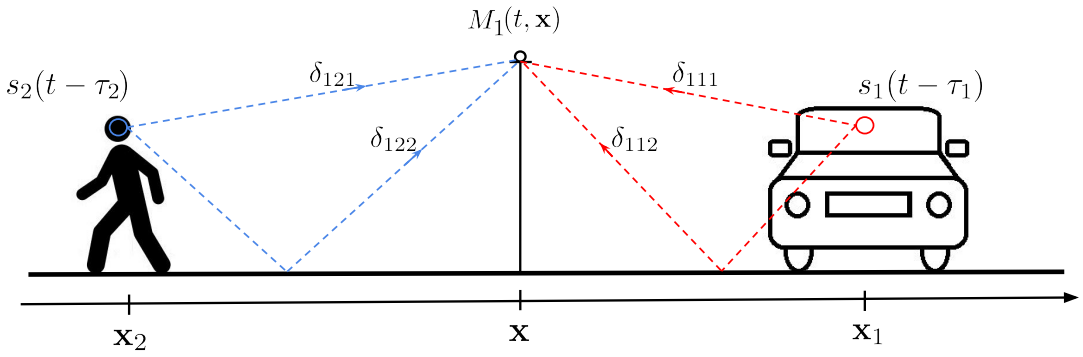


FIGURE 1.1 – Schéma du problème considéré en ESU pour un signal capté par un microphone au point $\mathbf{x}$. Deux sources sonores émises à l'instant $t - \tau_j$ à la position x_j sont présentes : une voiture, s_1 (résumée en une source ponctuelle symbolisée par un cercle rouge), et un piéton, s_2, (résumée en une source ponctuelle symbolisée par un cercle bleu). Chaque source se propage jusqu'au récepteur selon 2 chemins de propagation (champ direct δ_{ij1} et champ réfléchi δ_{ij2}).

présents dans l'environnement du récepteur et dont la somme globale s'exprime :

$$S_{tr.}(t) = \sum_{j=1}^{M} S_{v_j}(t), \tag{1.2a}$$

$$= \sum_{j=1}^{M} s_{v_j}(t - \tau_{v_j}) * \delta_j(t), \tag{1.2b}$$

où s_{v_j} correspond à l'émission sonore globale du véhicule j. Celle-ci a plusieurs origines : bruit du moteur thermique, aérodynamique et de roulement. Ici, puisque les citadins ne réalisent pas cette séparation mais considèrent l'ensemble de ces sources comme un tout, on résume la source *voiture* s_{v_j} comme l'ensemble de ses bruits. Précisons que le son émis par le klaxon du véhicule n'est pas considéré dans la source *voiture* car il appartient à la catégorie des avertisseurs sonores. L'ESU peut ainsi s'exprimer comme

$$M_i(t) = S_{tr.}(t) + S_{int.}(t) \tag{1.3}$$

où $S_{int.}(t)$ est la somme des autres sources sonores, appelées sources *interférantes*, qui n'appartiennent pas à la classe *trafic*. La description des émissions sonores des sources s'exprime également dans le domaine fréquentiel à l'aide d'une transformée de Fourier, telle que la source $S_j(t)$ s'exprime sous la forme :

$$TF\left[S_j(t)\right] = TF\left[s_j(t - \tau_j) * \delta_{ij}(t)\right], \tag{1.4a}$$

$$= \hat{s}_j(f)\hat{\delta}_{ij}(f)e^{-j2\pi f\tau_j}, \tag{1.4b}$$

$$= \hat{S}_j(f) \tag{1.4c}$$

avec la transformée de Fourier d'un signal $x(t)$:

$$TF\left[x(t)\right] = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi ft}dt, \tag{1.5a}$$

$$= \hat{x}(f) \tag{1.5b}$$

où f est la fréquence en Hz et j est, dans l'exponentielle, le nombre complexe. $\hat{s}_j(f)$ exprime l'émission sonore de la source j à la fréquence f, $\hat{\delta}_{ij}(f)$, la transformée de Fourier de la réponse impulsionnelle qui se traduit par un filtre de propagation. L'ESU s'exprime dans le domaine fréquentiel de la même façon que dans le domaine temporel (équation 1.3) :

$$\hat{M}_i(f) = \sum_{j=1}^{N} \hat{S}_j(f) \tag{1.6}$$

$$= \hat{S}_{tr.}(f) + \hat{S}_{int.}(f). \tag{1.7}$$

avec $\hat{S}_{tr.}(f) = TF\left[S_{tr.}(t)\right]$ et $\hat{S}_{int.}(f) = TF\left[S_{int.}(t)\right]$. $\hat{M}_i(f)$ équivaut alors à la contribution de chaque source sonore à la fréquence f au point i.

L'ESU $M_i(t)$ (et respectivement $\hat{M}_i(f)$) est donc le résultat de la somme d'un ensemble de sources sonores variées, ayant une allure temporelle, fréquentielle ainsi qu'un niveau sonore qui leurs sont propres, émises dans un environnement où des phénomènes de propagation complexes ont lieu. Les questions soulevées sont alors :

— **Comment caractériser les ESU ? Quels sont les moyens disponibles pour cela ?**

— **Comment estimer la présence et le niveau sonore du trafic routier ? Peut-on déterminer la contribution des autres sources sonores ?**

1.2 Utilisation de modèles prédictifs

L'une des premières pistes pour évaluer les ESU est l'utilisation de modèles prédictifs de bruits de trafic. L'objectif est alors de générer des lois d'émissions sonores en fréquences afin de déterminer les sources $\hat{s}_j(f)$ et des lois de propagation $\hat{\delta}_{ij}(f)$. De nombreux modèles d'émission existent afin de prédire les niveaux sonores émis par le trafic routier [Quartieri *et al.*, 2009], ferroviaire [van Leeuwen, 2000] et aérien [Zaporozhets et Tokarev, 1998]. Le trafic routier étant la source sonore la plus présente et la plus gênante dans l'environnement urbain, ce sont ces modèles qui sont ici présentés.

Plusieurs modèles visant à estimer la puissance acoustique, L_w, émise par les véhicules et la propagation des ondes sonores dans le milieu urbain ont été développés depuis plus de 30 ans, plusieurs pays ayant alors leur propre modèle (RLS-90 en Allemagne, CNR en Italie, NMPB-Routes en France, CoRTN au Royaume-Uni, Nord 2000 pour les pays Scandinaves). On se

propose tout d'abord de décrire succinctement 3 modèles selon plusieurs paramètres relatifs aux modèles d'émission et de propagation :

— Le modèle HARMONOISE/Imagine [Jonasson *et al.*, 2004], développé afin d'offrir un premier modèle d'émission harmonisé à l'échelle des pays membres de l'UE.

— La NMPB-routes-2008, un modèle français développé à partir des années 90 [SETRA, 2009a, SETRA, 2009b].

— Le modèle CNOSSOS-EU [CNO, 2012], le modèle développé le plus récent, là aussi afin d'harmoniser le modèle d'émission sonore et de propagation à l'échelle de l'UE, inspiré par les modèles précédents.

1.2.1 Modèle d'émission du trafic routier

Dans ces trois modèles, la puissance émise par une portion de route, L_w exprimée en dB, est liée à la puissance acoustique émise par un véhicule $L_{w,veh}$ et au débit des véhicules Q (nombre de véhicule/heure) :

$$L_w = L_{w,veh} + f(Q). \tag{1.8}$$

La puissance acoustique émise par le véhicule est décomposée en deux parties : une composante « bruit de roulement », L_{w_r}, et « bruit de moteur », L_{w_m} :

$$L_{w,veh} = 10 \times \log_{10}\left(10^{L_{w_r}/10} + 10^{L_{w_m}/10}\right). \tag{1.9}$$

Selon le modèle choisi, l'estimation de $L_{w,veh}$ est alors différente. Dans le Tableau 1.1, les niveaux de puissances L_{w_r} et L_{w_m} et la fonction $f(Q)$, exprimés en fonction de la vitesse v du véhicule en km/h, ainsi que plusieurs paramètres et catégories pris en considération, sont détaillés.

Les valeurs des coefficients dans les modèles HARMONOISE et CNOSSOS-EU, a_r , b_r , a_m et b_m, sont données selon les bandes de tiers d'octave et la catégorie du véhicule. Dans la NMPB, α_r, β_r, α_m et β_m sont données en fonction du type de véhicule, de sa vitesse et du revêtement du sol. Du Tableau 1.1, on observe que ces trois modèles, bien que similaires dans leur manière de décomposer la puissance acoustique d'un véhicule, présentent plusieurs aspects divergents : nombre de catégories de véhicules, nombre de revêtements des sols, nombre de bandes de tiers d'octave, calcul d'un niveau en dB pour le modèle HARMONOISE et CNOSSOS-EU et d'un calcul en dB par mètre pour le modèle NMPB. Ces 3 modèles considèrent également des vitesses stables, en accélération ou en décélération au travers de corrections qui cependant diffèrent entre les modèles. Si les modèles HARMONOISE et CNOSSOS-EU sont similaires sur l'estimation des niveaux de puissance L_{w_r} et L_{w_m}, le premier considère deux sources ponctuelles pour modéliser un véhicule alors que le second n'en considère qu'une. De plus, dans ces deux modèles, si le nombre de catégories de véhicules est identique, leur classification est différente

Tableau 1.1 – Paramètres d'estimation de la puissance acoustique selon 3 modèles d'émission sonore.

	Harmonoise	NMPB	Cnossos-Eu
$L_{w,veh}$	$L_{w_r} = a_r + b_r \log\left(\frac{v}{70}\right)$ $L_{w_m} =$ $a_m + b_m \left(\frac{v-70}{70}\right)$	$L_{w_r} =$ $\alpha_r + \beta_r \log\left(\frac{v}{90}\right)$ $L_{w_m} =$ $\alpha_m + \beta_m \log\left(\frac{v}{90}\right)$	$L_{w_r} = a_r + b_r \log\left(\frac{v}{70}\right)$ $L_{w_m} =$ $a_m + b_m \left(\frac{v-70}{70}\right)$
$f(Q)$	$\log\left(\frac{Q}{1000 \times v}\right)$	$\log(Q)$	$\log\left(\frac{Q}{1000 \times v}\right)$
bandes de fréquences	25 Hz - 10 kHz	100 Hz - 5 kHz	125 Hz - 4 kHz
description de la source	2 sources équivalentes ponctuelles	1 source équivalente ponctuelle	1 source équivalente ponctuelle
nombre de catégories de véhicules	5	2	5
nombre de revêtements	1	3	1

puisque Harmonoise considère une 5e catégorie pour les camions de type tracteur, camion de chantier alors que celle de Cnossos-Eu permet d'inclure les véhicules électriques.

1.2.2 Modèle de propagation

Chaque méthode possède également son modèle de propagation acoustique qui simule le rayonnement acoustique de la source. C'est cette étape qui est la plus longue à calculer en raison du volume du domaine Ω à considérer, du nombre de chemins de propagation k possible et du nombre M de sources sonores à prendre en compte. Dans le cas du modèle d'Harmonoise, le calcul de la propagation du son est réalisée en considérant plusieurs approches (résolution de l'équation parabolique, méthode des éléments de frontières ou tir de rayons) afin de s'adapter à différentes configurations, en y considérant des conditions atmosphériques homogènes. L'influence de la route est prise en compte selon son revêtement (température, âge, humidité). Cette approche est reconnue pour être plus « physique » mais ausi pour avoir un temps de calcul plus long que les autres modèles (jusqu'à 50 fois plus long) [Probst *et al.*, 2011]. Pour la NMPB, la méthode de propagation choisie est celle des tirs de rayons où les chemins directs, réfléchis et diffractés sont considérés entre la source et le récepteur. En fonction des conditions atmosphériques relevées (température, vent), les atténuations dans les conditions favorables (à la propagation) et homogènes sont considérées. Les effets de sol (3 types de routes considérés, dont les propriétés varient selon leur ancienneté et la température ambiante de l'air), la divergence géométrique et l'absorption atmosphérique sont aussi pris en compte. Enfin, les aspects liés à la propagation du son du modèle NMPB ont été repris dans le modèle Cnossos-Eu.
Là encore, les modèles de propagation diffèrent selon les méthodes choisies. Une comparaison

plus détaillée entre plusieurs de ces modèles, et sur de nombreux autres points, peut être trouvée dans [Steele, 2001] et dans [Garg et Maji, 2014].

1.2.3 Réaliser des cartes du bruit de trafic en ville

À partir de 2002 est instaurée la directive européenne 2002/49/CE *relative à l'évaluation et à la gestion du bruit dans l'environnement* [Parlement Européen, 2002] dont le but est de mieux connaitre la répartition des niveaux sonores générés par les 4 sources sonores les plus bruyantes (le trafic routier, ferroviaire et aérien ainsi que les Installations Classées pour la Protection de l'Environnement (ICPE)) dans les agglomérations de plus de 100 000 habitants. Cette directive prévoit :

— d'évaluer l'exposition au bruit des populations en se basant sur des méthodes communes aux pays européens,

— d'informer les populations sur leur niveau sonore d'exposition et sur les effets du bruit sur la santé,

— de connaitre et de délimiter les zones bruyantes et les zones calmes.

Cette directive se traduit notamment par la production de cartes de bruits stratégiques, pour chacune de ces 4 sources sonores, afin de déterminer les endroits où les niveaux sonores sont élevés. Un résumé des étapes permettant la réalisation des cartes de bruit du trafic routier est présenté en Figure 1.2.

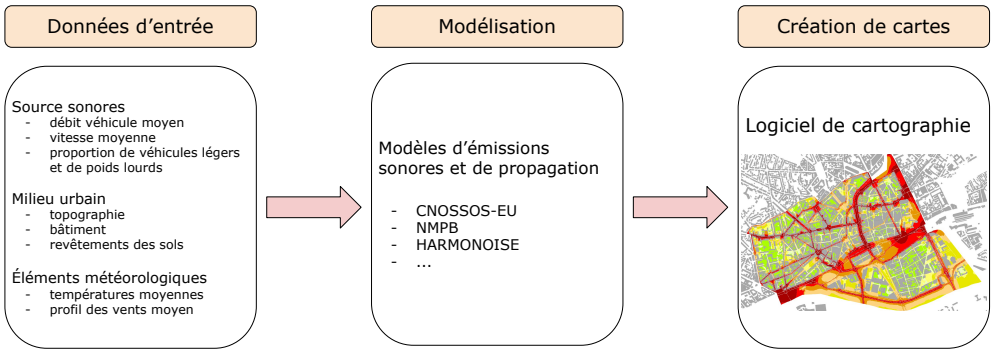


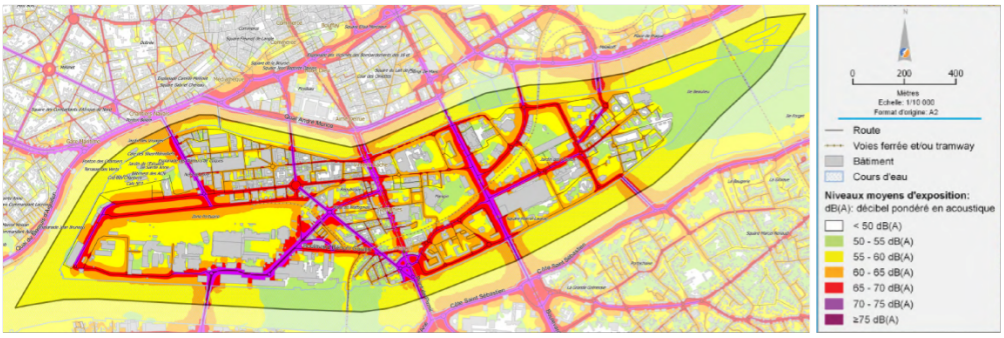
FIGURE 1.2 – Résumé des étapes menant aux cartes de bruit du trafic routier.

Dans un premier temps, plusieurs indicateurs en données d'entrée des modèles d'émission sonore sont relevés *in situ* :

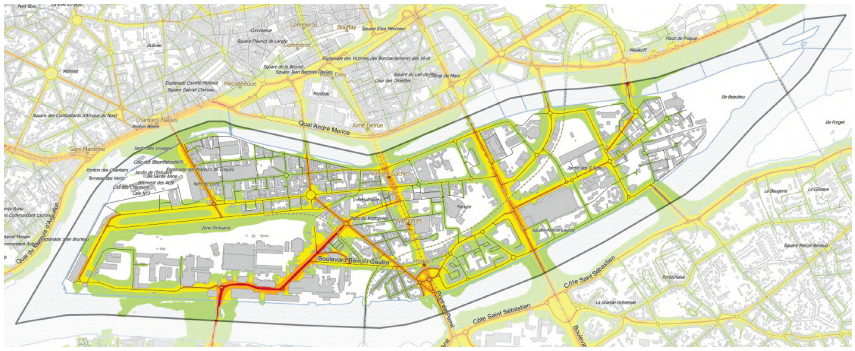
— vitesses moyennes des véhicules sur les portions de routes principales,

— débits de véhicules (nombre de véhicules par tranche horaire),

— composition du trafic (nombre de véhicules légers et de poids lourds),



(a)



(b)

FIGURE 1.3 – L_{DEN} (a) et L_N (b) de l'île de Nantes pour le trafic routier [Nantes Métropole, 2017].

— architecture et topographie de la ville (revêtement au sol),

— conditions météorologiques (températures, vent).

Puis à partir d'un modèle d'émission sonore et de ces données d'entrée, il est possible d'établir la puissance émise par chaque véhicule pour chaque fréquence, $\hat{s}_{v_j}(f)$. Le temps de calcul associé à la modélisation des sources sonores est quasi instantané. Enfin, l'influence de l'environnement $\hat{\delta}(f)$ sur la répartition des niveaux sonores dans la ville est calculé à l'aide d'un logiciel numérique (Mithra, Immi, CadnAa...). Les logiciels accompagnés d'un Système d'Information Géographique (SIG) sont ceux qui offrent actuellement le plus de possibilités. Un SIG est un outil informatique conçu pour stocker, analyser et manipuler plusieurs types de données spatiales et géographiques comme l'architecture des villes ou le nombre d'habitants présents. Son utilisation permet de connaitre plus facilement le nombre de citadins exposés à des forts niveaux sonores. Par exemple, le logiciel OrbisGIS [1], destiné à représenter des données spatiales, permet la réalisation de cartes de bruit par l'ajout d'un plugin, *NoiseModelling*, développé par [Fortin *et al.*, 2012].

1. `http://orbisgis.org/`

Les cartes de bruits produites résument les niveaux sonores équivalent pondérés A sur 24h (L_{DEN} pour *Day-Evening-Night* (*Jour-Soir-Nuit* en français)) et durant la nuit (L_N) :

$$L_{DEN} = 10 \times \log_{10}\left(\frac{1}{24}\left(12 \times 10^{\frac{L_D}{10}} + 4 \times 10^{\frac{L_E+5}{10}} + 8 \times 10^{\frac{L_N+10}{10}}\right)\right) \tag{1.10}$$

avec L_D, L_E et L_N, les niveaux sonores équivalent pondéré A pour les périodes respectives 6h-18h, 18h-22h, 22h-6h (horaires pouvant être changées suivant le rythme de vie des habitants de la zone considérée),

$$L_D = 10 \times \log_{10}\left(\frac{1}{T}\sum_{t=1}^{T} 10^{\frac{L_{D_t}}{10}}\right), \tag{1.11a}$$

$$L_E = 10 \times \log_{10}\left(\frac{1}{T}\sum_{t=1}^{T} 10^{\frac{L_{E_t}}{10}}\right), \tag{1.11b}$$

$$L_N = 10 \times \log_{10}\left(\frac{1}{T}\sum_{t=1}^{T} 10^{\frac{L_{N_t}}{10}}\right), \tag{1.11c}$$

avec L_{X_t}, le niveau sonore dans la tranche horaire t. Les niveaux L_E et L_N sont majorés respectivement de 5 dB(A) et de 10 dB(A) afin de pénaliser les plages horaires où la gêne occasionnée par le trafic est plus importante. La Figure 1.3 résume, par exemple, le L_{DEN} et le L_N pour le bruit du trafic routier dans un quartier de la ville de Nantes. La réalisation de ces cartes permet ensuite d'estimer le nombre de personnes touchées par des niveaux sonores trop élevés selon la réglementation en vigueur et facilite ainsi la mise en oeuvre de travaux d'aménagement (construction de mur anti-bruit, changement de revêtement, diminution des vitesses...). L'utilisation de modèles prédictifs présente alors l'intérêt de pouvoir simuler l'effet de ces dispositifs sur les niveaux sonores émis et d'en mesurer l'impact [Murphy et King, 2011, Guedes *et al.*, 2011]. Les cartes produites doivent ensuite être mises à jour tous les 5 ans.

1.2.4 Vers la modélisation d'autres sources sonores ?

Le trafic routier, aérien et ferroviaire et les ICPE sont actuellement les seules sources sonores à faire l'objet de cartes de bruit. Mais il est tout à fait possible d'ouvrir ces outils à d'autres sources sonores. En effet, les cartes de bruits, si elles permettent de mieux identifier la présence de bruits en ville, ne permettent pas de représenter au mieux la perception des citadins de l'ESU [Brown, 2012]. À travers plusieurs études perceptives [Lavandier et Defréville, 2006, Hong et Jeon, 2013], plusieurs sources sonores comme la voix, le chant d'oiseaux, le bruit des fontaines, ont montré avoir une influence dans la perception de l'ESU. Il peut donc être intéressant et utile de savoir estimer leur présence dans les villes afin de permettre une représentation des ESU à partir d'indicateurs physiques non plus juste de sources sonores connotées négativement (comme le trafic routier) mais de sources plus appréciées. Certaines ont déjà fait l'objet de modélisation (dans le cadre d'autres études ou applications) comme les oiseaux [Nemeth *et al.*, 2013], les voix [Hayne *et al.*, 2011] ou les fontaines [Watts *et al.*, 2009], mais ne sont, pour l'instant, pas assez étudiées pour offrir des modèles validés. Une autre difficulté est la localisation de certaines de

ces sources dans l'espace urbain. Si les sources sonores fixes comme les fontaines ou les cloches d'une église sont faciles à localiser, d'autres, comme la foule et les oiseaux, sont plus difficiles à déterminer car plus mobiles et parcimonieuses. Dans [Aumond *et al.*, 2018b], c'est par une approche statistique que la position des sources est déterminée : les piétons sont ainsi plus susceptibles de se trouver sur des places ou le long des trottoirs, alors que pour les oiseaux, ce sont dans les parcs ou auprès des arbres qu'on les trouve. Il deviendrait alors envisageable de générer des cartes dites multi-sources d'une ville où une représentation des niveaux sonores émis par chaque source pourrait être générée et ainsi de s'orienter vers des cartes de bruits perceptives.

1.2.5 Limitations des modèles prédictifs

L'utilisation de modèles d'émissions sonores présente donc plusieurs intérêts : prédiction des niveaux sonores dans une situation donnée, création de cartes de bruit de trafic et possibilité de tester différents scénarios d'aménagements. Leur utilisation s'est répandue et démocratisée en raison de l'introduction de la directive européenne. Les premières études réalisées suivant les recommandations de cette directive ont toutefois permis de soulever plusieurs limites à ces méthodes comme celles liées au choix du modèle parmi ceux existants ou aux incertitudes liées aux estimations des données d'entrées.

1.2.5.1 Limites liées à la modélisation de la source et de la propagation

Ces modèles sont basés sur des mesures et des études physiques du trafic. Comme tout modèle simulant la réponse d'un système physique, il n'est pas possible d'en générer un universel qui puisse s'adapter à l'ensemble des scénarios possibles. Des simplifications sont ainsi réalisées, par exemple en classifiant les routes et le parc automobile en un nombre réduit de catégories ou bien en considérant (ou non), en plus des conditions atmosphériques homogènes, des conditions favorables à la propagation. Ce sont d'autant de simplifications qui, certes, facilitent l'implémentation et l'utilisation des modèles mais qui sont aussi vecteurs d'incertitudes. De plus, ces catégorisations prennent le risque de mal prendre en compte certains cas limites qui ne correspondent pas spécifiquement à ceux définis.

Un second aspect, évoqué dans les parties 1.2.1 et 1.2.2 est celui de l'existence de plusieurs modèles d'émission et de propagation au sein de plusieurs pays européens. Avant même la mise en place de la directive, [Steele, 2001] en avait comparé plusieurs selon différents aspects (données d'entrée, type de cartographie, méthode de propagation des différents logiciels). Parmi ces différents outils, l'auteur met en avant le problème, soulevé également par [King *et al.*, 2011], de la diversité des méthodes de calculs qui peuvent être employées : quelle méthode, parmi celles existantes, doit-être utilisée ? Dans un premier temps, ce choix a été laissé libre par la directive européenne. Les premières cartes de bruits ont donc été établies sur des modèles différents : par exemple pour la même année, dans [Kliučininkas et Šaliūnas, 2006], le modèle RLS-90 est employé pour calculer la carte de bruit dans le centre-ville de Kaunas, en Lituanie, alors que dans [Murphy *et al.*, 2006], la carte de bruit de trafic dans la ville de Dublin,

en Irlande, est construite sur la base du modèle HARMONOISE. Une comparaison exhaustive de 8 modèles (FHWA, CoRTN, RLS-90, ASJ, HARMONOISE/Imagine, Son Road, Nord 2000 et NMPB-Routes-2008) a également été réalisée par Garg et Maji [Garg et Maji, 2014] selon un plus grand nombre de critères (modélisation des sources sonores, vitesse des véhicules (constantes, accélération/décélération, intersection...), modèle de propagation, modélisation des effets de sol, effets météorologiques...). À travers leur comparaison, les auteurs relèvent ainsi les nombreuses différences notamment entre les modèles de propagation du son. Les auteurs de l'étude précisent tout de même qu'il est difficile de déterminer un « meilleur » modèle par rapport aux autres, chacun ayant ses avantages et ses limites. Afin de résoudre ces problèmes, la méthode CNOSSOS-EU [CNO, 2012, Kephalopoulos *et al.*, 2012] a ainsi été développée, basée sur les méthodes déjà existantes en vue d'harmoniser la construction des cartes de bruit des villes à l'échelle européenne pour faciliter leur comparaison.

Toutefois, quel que soit le modèle choisi, la confrontation des niveaux sonores prédits face à des mesures faites en ville reste à réaliser même si l'ensemble de ces modèles d'émission et de propagation a été développé et validé à partir de mesures faites dans des conditions optimales. Mais, la comparaison entre les niveaux sonores calculés et mesurés reste délicate. Premièrement, les mesures présentent l'inconvénient d'être soumises à d'autres sources sonores qui ne sont pas liées au trafic et qui viennent donc fausser les estimations des niveaux sonores. De plus, il faut s'assurer, lors des mesures, que les données d'entrée des modèles correspondent bien aux conditions expérimentales, afin de comparer correctement les mesures et les estimations simulées, ce qui n'est pas facile.

Enfin, ces modèles dépendent des données d'entrée relevées *in situ* s'exprimant sous la forme de moyennes et qui induisent donc des écarts-types qui se propagent dans les étapes suivantes du calcul. [Van Leeuwen et Van Banda, 2015] proposent un résumé et un schéma détaillé de la propagation de ces erreurs sur l'ensemble du modèle.

1.2.5.2 Limites liées à la simulation et à la représentation

La réalisation de cartes de bruit est une opération qui peut être très lourde en coût et en temps de calculs (de quelques heures à plusieurs jours). Dans le cas de la cartographie de bruit de trafic, les modèles de sources $s_j(t)$ et de propagation $\delta_{ij}(t)$ sont implémentés dans des logiciels numériques pour déterminer, sur l'ensemble d'une ville ou d'un quartier, leurs niveaux sonores. Cette numérisation implique alors la discrétisation de l'environnement urbain qui consiste à décrire l'espace urbain Ω en un ensemble de points qui forme un maillage. L'environnement urbain est alors réduit en un espace discret Ω_m. Le plus souvent, c'est un maillage régulier de 10 mètres par 10 mètres qui est choisi. À l'échelle d'une ville, c'est ainsi plusieurs millions de points qui peuvent être définis. Chaque source est alors rattachée à un point du maillage ($s_j(t)$ pour $j \in \Omega_m$). Le calcul consiste ensuite à estimer la propagation acoustique entre les sources et les points de ce maillage, selon les différents chemins de propagation possibles (δ_{ij,Ω_m}), pour chaque bande de fréquences. Plus le nombre de points dans Ω_m est élevé, plus le temps nécessaire pour calculer le niveau sonore augmente. Également, le nombre de chemins de propagation d'une onde

acoustique est également une variable que contrôle l'utilisateur. D'une valeur théorique infinie, celle-ci se réduit à un nombre K qui influe alors directement sur la quantité d'énergie finale présente en un point i. L'équation 1.1c devient alors :

$$M_{i,\Omega_m}(t) = \sum_{j=1}^{N} \sum_{k=1}^{K} s_j(t - \tau_{ijk})\delta_{ijk} \tag{1.12}$$

avec $\{i,\ j\} \in \Omega_m$. Le choix de ces paramètres implique donc un compromis à faire, par l'utilisateur, entre la précision des résultats souhaitée et le temps de calcul alloué. L'étape de discrétisation du milieu, si elle simplifie la forme de l'environnement urbain, présente l'inconvénient de ne pas prendre en compte les multiples variations géométriques des façades ou bien l'ensemble des petits mobiliers urbains qui ont un rôle dans la diffusion du son. De plus, en raison de la discrétisation du milieu, des méthodes d'interpolation (méthode linéaire, de krigeage) sont utilisées pour calculer les niveaux sonores entre ces points, ce qui est également source d'approximation pouvant mener à de mauvaises interprétations [Van Leeuwen et Van Banda, 2015].

Enfin, une fois la carte générée, seulement 2 niveaux sonores par source de trafic sont obtenus, L_{DEN} et L_N, et mis à jour tous les 5 ans. C'est donc une information statique et restreinte qui est obtenue. Cependant, le trafic routier, ferroviaire et aérien varient aussi bien à l'échelle de l'année, d'une journée ou même d'une heure [Lv *et al.*, 2015]. S'il parait envisageable de calculer des cartes de bruit pour, par exemple, chaque heure de la journée, cela reste une opération très longue et coûteuse à réaliser. L'utilisation de modèles dynamiques de trafic, couplés aux modèles d'émissions sonores [Can *et al.*, 2010], n'est actuellement pas destinée à la cartographie des ESU mais est une piste envisageable pour modéliser l'impact de l'écoulement du trafic routier à l'échelle d'une rue ou d'une intersection.

En conclusion, l'utilisation de modèles prédictifs est une approche utile pour réaliser une première estimation de la répartition du bruit de trafic en ville. Si elle présente des avantages (estimation d'un niveau physique à l'échelle de la ville, possibilité de tester des scénarios d'aménagement), elle présente plusieurs limites :

— le nombres de type de sources est, pour l'instant, trop restreint (trafic routier, aérien, ferroviaire, ICPE) et ne permet pas de considérer l'ensemble des ESU et la perception qu'en ont les citadins,

— l'utilisation d'un modèle d'émission et de propagation et le calcul numérique entraine de nombreuses incertitudes qui sont difficiles à estimer,

— l'accès à l'évolution temporelle des sources sonores n'est pas possible.

Afin de considérer l'ensemble des sources et des évènements sonores présents en ville et de compléter les estimations générées par les modèles prédictifs, une autre approche est envisagée, basée sur la réalisation de mesures et d'enregistrements sonores réalisés directement dans la ville.

1.3 Utilisation de mesures acoustiques

À la différence des modèles prédictifs qui déterminent l'émission sonore des sources $s_j(t)$ et leur propagation $\delta_{ij}(t)$ pour en déterminer le niveau sonore en un point i, la réalisation de mesures donne directement accès au mélange global $M_i(t)$. Les limites liées à la modélisation des sources, de la propagation du son et de l'ensemble des environnements urbains dans toute leur complexité n'interviennent alors plus dans les mesures faites *in situ*. Plusieurs études ont déjà été menées en ville en faisant intervenir des mesures pour des durées plus ou moins longues (de quelques jours [Romeu *et al.*, 2011] à plusieurs années [Gaja *et al.*, 2003]) avec des microphones de mesures de haute qualité. Dans une première étude [Zannin *et al.*, 2002] a réalisé une série de mesures sur près de 1000 positions dans la ville de Curitiba, au Brésil, pour étudier l'ESU dans sa globalité afin d'évaluer l'exposition des citadins aux bruits dans les différents quartiers de la ville, puis a réduit la surface d'étude dans [Zannin *et al.*, 2013] où 58 points de mesures sont déployés dans le campus universitaire de la ville. Dans [Mioduszewski *et al.*, 2011], 40 microphones sont placés isolément à travers la ville de Gdansk en Pologne pour une durée d'un an afin d'y mesurer le niveau sonore du trafic et de valider la cartographie de bruit.

Il y est donc tout à fait possible d'étudier les ESU et les sources qui les composent à l'aide d'enregistrements et de mesures acoustiques faites directement dans la ville. Plusieurs approches sont possibles aux travers de mesures faites par des réseaux de capteurs, par des mesures mobiles et participatives, chacune ayant déjà fait l'objet d'études et d'applications. Celles-ci sont successivement présentées pour ensuite être interrogées au regard de notre problématique.

1.3.1 Déploiement de réseaux de capteurs fixes

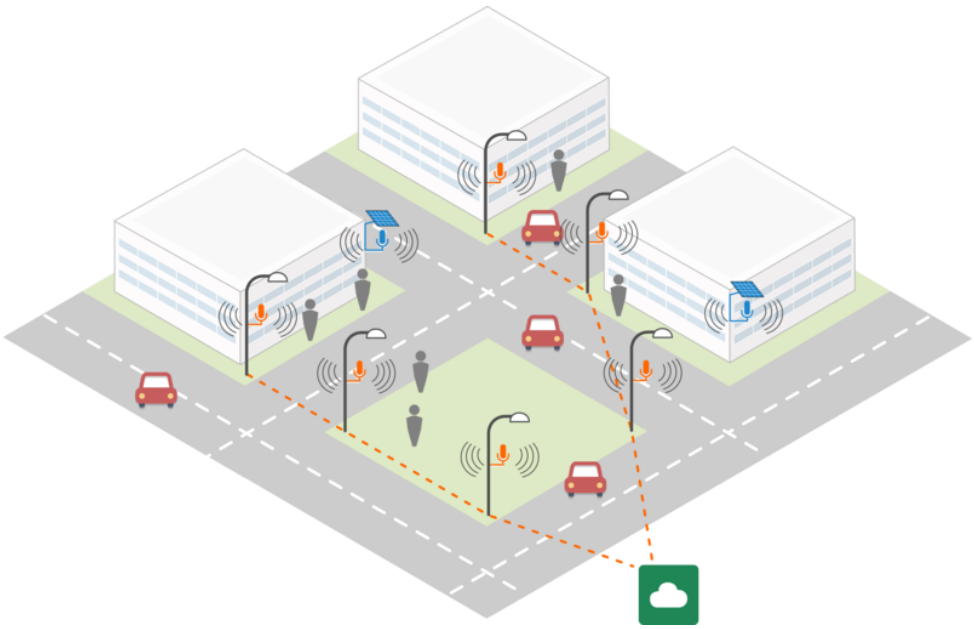


FIGURE 1.4 – Schéma d'un réseau de capteurs fixes. [2]

Les approches évoquées en introduction de cette partie ont consisté en des mesures limitées dans le temps, or la pérennisation de l'installation de capteurs en ville est de plus en plus envisagée à l'heure de l'émergence de l'*IoT* (*Internet of Things* ou l'internet des objets) [Zanella *et al.*, 2014] et de la ville intelligente (*Smart City*) [Chourabi *et al.*, 2012]. De nombreuses villes s'équipent actuellement en réseaux de différents types de capteurs disséminés dans le milieu urbain afin d'en contrôler, en temps réel, de nombreux aspects : distribution d'énergie, gestion des transports, de l'eau ou des déchets. L'objectif étant alors d'optimiser le fonctionnement de la ville afin d'améliorer la qualité de vie des citadins. Ces réseaux sont constitués d'un ensemble de capteurs fixes alimentés et connectés à un réseau de télécommunication. Leurs mesures sont alors stockées par des serveurs pour ensuite être post-traitées. Dans le domaine de l'acoustique, l'intérêt principal d'un tel réseau est la possibilité d'avoir accès à des variations à long-terme des niveaux sonores à l'emplacement des microphones. Le schéma d'un tel réseau est résumé en Figure 1.4. Un premier réseau de capteurs développé depuis plusieurs années est celui de la ville de Paris, géré par BruitParif, à travers le projet RUMEUR [Mietlicki *et al.*, 2012], où des microphones sont déployés afin d'évaluer l'ESU en région parisienne. Le réseau comprend des stations fixes mais intègre aussi des campagnes de mesures plus courtes (de plusieurs heures à plusieurs années) pour voir l'impact acoustique d'un évènement (« Journée sans voiture » par exemple) ou d'un aménagement (modifications de voirie). Un site internet [3] est mis à disposition pour avoir un aperçu complet des mesures réalisées sur les nombreux emplacements. En parallèle, plusieurs études se sont intéressées à la prise en compte de mesures pour caractériser le bruit du trafic comme dans [Makarewicz et Galuszka, 2011] qui propose une estimation des niveaux sonores L_{DEN} et L_N directement à partir des mesures et dont l'incertitude est estimée en fonction de la durée d'acquisition des mesures. Dans [Wei *et al.*, 2016], une technique d'assimilation de données est proposée afin de réaliser des cartes de bruit dynamiques basées sur des mesures. Les niveaux de puissances et les effets de propagation de chaque source sont ici modifiés par des termes correctifs obtenus en minimisant l'erreur quadratique entre les niveaux prédits et les niveaux mesurés. L'arrivée sur le marché, depuis plusieurs années, de capteurs acoustiques à bas coût (moins de 10 € pour un capteur microphone MEMS) [Van Renterghem *et al.*, 2010] rend possible le déploiement d'un plus grand nombre de capteurs acoustiques à travers les villes. La mise en place de tels réseaux donne lieu à plusieurs projets où différentes questions sont étudiées (nombre de capteurs installés, surface couverte par ce réseau, mesures fixes ou mobiles)...comme le projet européen DYNAMAP [4] [Xavier *et al.*, 2016]. DYNAMAP a pour objectif de développer un système de cartographie de bruit dynamique basé sur des réseaux de capteurs à bas coûts installés en ville. Une application de ce projet a déjà été réalisée dans deux villes tests, Milan et Rome [Bellucci et Zambon, 2017]. Le principe de leur approche est d'ajuster les cartes de bruits simulées en tenant compte des différences obtenues entre les niveaux sonores mesurés aux stations et les niveaux sonores calculés à ce même point par les modèles prédictifs. Pour limiter le coût d'un tel déploiement, le nombre de microphones est réduit en les installant à

2. `http://cense.ifsttar.fr/`
3. `http://rumeur.bruitparif.fr/`
4. `http://www.life-dynamap.eu/`

des emplacements spécifiques représentatifs des différents scénarios possibles de trafic routier (homogénéité du trafic, type de revêtement) [Zambon *et al.*, 2017]. Le projet SONYC [5] à New-York dédie son réseau de capteurs à la surveillance de la pollution sonore et au développement d'outils de traitement du signal afin de décrire l'ESU par l'étude des sources présentes [Mydlarz *et al.*, 2017]. Enfin, le projet CENSE [6] vise à développer un réseau de capteurs dans la ville test de Lorient afin là encore d'améliorer la cartographie du bruit de trafic en agrégeant les données simulées des niveaux sonores du trafic avec les mesures réalisées en ville par ce réseau. L'approche est différente de DYNAMAP, puisqu'ici l'étude se restreint à l'échelle de plusieurs quartiers de la ville afin d'avoir un réseau de capteurs dense. La mise à jour des cartes est faite à l'aide de techniques d'assimilation de données en vue de compléter les cartes de bruits prédites avec les mesures. Ces méthodes d'assimiliation sont notamment utilisées dans le domaine des sciences géophysiques et consistent à modifier une estimation émises par un modèle prédictif à partir de données mesurées [Wu *et al.*, 2008]. Le projet s'intéresse également à la perception des citadins des ESU par de questionnaires qui leur sont envoyés et de mesures réalisées par ce réseau de capteurs.

Toutefois, l'installation de tels réseaux de capteurs nécessite de gérer de nombreuses problèmes techniques comme la disposition des microphones, leur maintenance, la transmission et le stockage des mesures, leur alimentation électrique... Une des premières limites techniques est la performance individuelle des capteurs. En effet, la réduction de leur coût s'est faite en diminuant leur performance individuelle (dynamique énergétique et fréquentielle) [Mydlarz *et al.*, 2015]. Il est donc nécessaire de caractériser chacun des microphones en vue de connaitre leurs performances et leur limites. Un des points cruciaux est celui de la position des microphones des mesures (quelle hauteur ? quelle position par rapport à des sources sonores qui seraient dignes d'intérêt ?) et de la surface couverte par ces mesures. Un réseau distribué selon un maillage dense permettra une bonne représentation de l'espace mais coûtera cher à installer et à maintenir alors qu'une faible densité de capteurs sera moins onéreuse mais apportera moins d'information et nécessitera des interpolations entre les mesures, ce qui reste une source d'incertitudes. Toutefois, la réalisation de mesures acoustique en ville n'est pas nécessairement obligée d'être réalisée via des réseaux de capteurs fixes. D'autres pistes sont également explorées.

1.3.2 Mesures mobiles

En parallèle aux réseaux fixes, la mesure mobile est une voie envisagée. Elle consiste à réaliser des mesures acoustiques en plaçant le microphone sur un support mobile (piéton, cycliste, voiture, bus). Ce type de mesure correspond aux mesures faites par des professionnels avec du matériel de haute qualité.

Dans le cadre des études des ESU et du bruit de trafic, l'avantage de cette méthode par rapports aux capteurs fixes est sa capacité à pouvoir couvrir plus facilement une plus grande surface urbaine à moindre coût. Les mesures mobiles sous-entendent deux manières d'être réalisées : soit

5. `https://wp.nyu.edu/sonyc/`
6. `http://cense.ifsttar.fr/`

le microphone réalise sa mesure sur un support mobile qui se déplace en même temps [Alsina-Pagès *et al.*, 2016]. Dans ce cas, un traitement du signal doit être effectué pour prendre en compte le bruit émis par ce support. Le microphone peut également être placé sur un support mobile afin de le déplacer pour faire ensuite des mesures fixes [Manvell *et al.*, 2004] ce qui simplifie la tâche mais nécessite plus de temps pour couvrir une surface similaire par rapport aux mesures faites sur un support mobile. L'inconvénient de ces méthodes est qu'elles ne permettent pas la réalisation de mesures à long terme et donc d'estimer l'évolution temporelle des niveaux sonores en un point donné au cours du temps. Ainsi, plusieurs travaux se sont intéressés à l'agrégation des mesures mobiles à des mesures réalisées par des stations fixes. [Morillas et Gajardo, 2014] s'intéressent aux incertitudes sur l'estimation des niveaux sonores estimés suivant le nombre de points ou le nombre de jours de mesures. Dans [Can *et al.*, 2014], la prise en compte de mesures mobiles pour compléter des stations fixes est comparée à des méthodes d'interpolation (méthode de Kriging, pondération inverse de la distance). Il en résulte que l'apport des mesures mobiles diminue l'erreur produite par rapport aux méthodes d'interpolation en cela qu'elles permettent d'apporter plus d'informations quant aux variations spatiales du niveau sonore (rues calmes peu fréquentées, rues très passantes, aux abords d'intersections...), ce que ne permet pas une méthode d'interpolation numérique [Aumond *et al.*, 2018a].

La réalisation de mesures mobiles est notamment très courante pour des études perceptives de l'ESU par des citadins lors de *soundwalks* (*marches sonores* en français). Cette méthode consiste à réaliser un parcours en ville et à soumettre, à un panel d'auditeur, un questionnaire sur les sons qui les entourent. L'enregistrement de l'ESU durant cette marche permet alors de corréler leurs réponses et leur perception avec des indicateurs (niveau sonore en dB(A), présence de certaines sources sonores...) extraits de ces enregistrements [Brocolini *et al.*, 2013a, Hong et Jeon, 2013]). Dans [Aumond *et al.*, 2017a], l'évaluation perceptive d'ESU par des citadins est corrélée à des indicateurs physiques tels que le niveau sonore fractile L_{50} dans la bande de tiers d'octave de 1 kHz ainsi que la variation normalisée en temps et en fréquence des bandes de 500 Hz et de 4 kHz. Enfin, les mesures mobiles peuvent servir à des études plus ponctuelles par exemple en vue de classer les ESU selon différents indicateurs physiques comme dans [Rychtáriková et Vermeir, 2013], où 370 enregistrements de 15 à 20 minutes réalisés dans 4 villes de Belgique ont été réalisés lors de *marches sonores*, et dans [Can et Gauvreau, 2015], où des enregistrements mobiles ont été réalisés dans la ville de Marseille.

1.3.3 Mesures participatives

Une dernière voie sollicite la participation des citadins. Ces mesures participatives peuvent se réaliser en équipant les citadins de dispositifs spécifiques [Delaitre *et al.*, 2014] ou bien à partir d'applications développées pour smartphones qu'ils peuvent télécharger eux-mêmes. Profitant de la démocratisation de ces appareils et de l'augmentation de leurs performances, ces applications leurs permettent d'avoir un dispositif suffisamment performant pour mesurer les niveaux sonores autour d'eux. Cette approche permet surtout d'obtenir un plus grand nombre de mesures qui ont le plus souvent une distribution spatiale et temporelle plus aléatoire mais

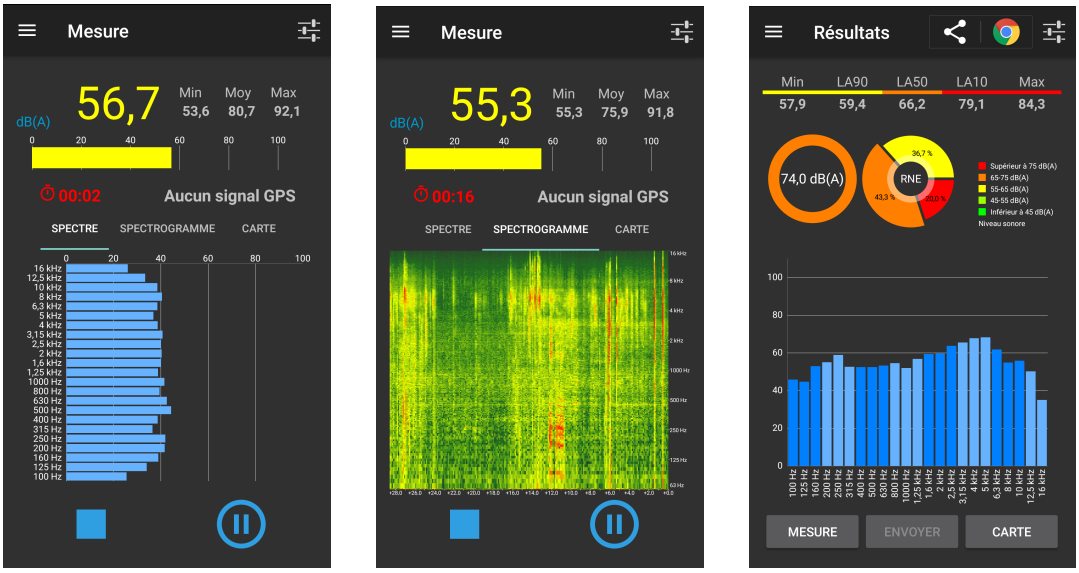


FIGURE 1.5 – Captures d'écran de l'application *NoiseCapture*.

qui sont aussi effectuées moins régulièrement. L'utilisation de ces mesures est toutefois encore sujette à caution puisque de nombreux problèmes sont encore à résoudre comme la calibration et la prise en compte des performances des microphones dans les faibles et forts niveaux sonores [Aumond *et al.*, 2017b] ou bien encore la qualité de la réalisation de la mesure faite par l'utilisateur... Dans ce cas, le traitement statistique des résultats est primordial afin de détecter les mesures incongrues pour ne pas les considérer [Guillaume *et al.*, 2016]. Plusieurs applications ont été développées comme *NoiseSpy* [Kanjo, 2010] ou *Ambicity* [Ventura *et al.*, 2017]. On peut également relever dans le projet *Noise Planet*[7] l'application pour smartphone, *NoiseCapture* [Guillaume *et al.*, 2016] (Figure 1.5), qui permet, là aussi, à l'utilisateur d'évaluer les niveaux sonores l'entourant tout en ayant la possibilité de décrire, à l'aide de mots-clés prédéfinis, les sons présents et l'ambiance sonore de la scène. La géo-localisation et les mesures sont ensuite collectées puis traitées pour produire des cartes de bruits, publiées en ligne (voir Figure 1.6). Les mesures réalisées par ces dispositifs permettent d'évaluer les ESU toutes sources confondues sans considérer leur influence individuelle. En outre, un des intérêts de ces applications, en plus de sensibiliser le citadin à son environnement sonore, est de le rendre producteur et utilisateur de données environnementales. Les informations récoltées sur ces trajets permettent de calculer son exposition au bruit ou bien de le guider vers des itinéraires secondaires où son exposition au bruit serait plus faible [Aumond *et al.*, 2016].

7. `http://noise-planet.org`

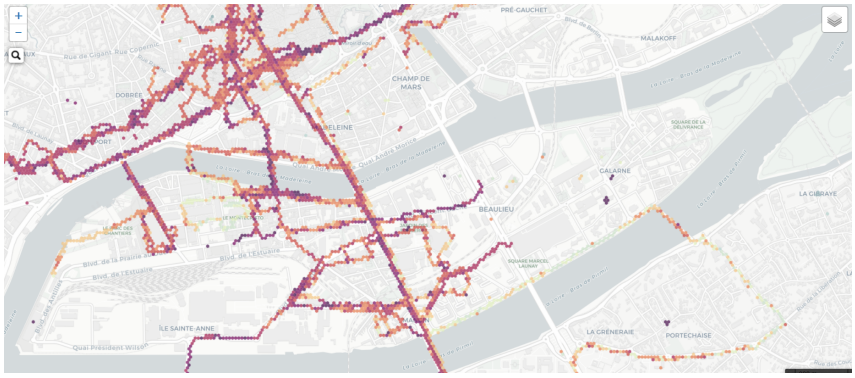


FIGURE 1.6 – Carte de l'ESU de l'île de Nantes mesurée par l'application *NoiseCapture* (relevée le 22/03/2018).

1.3.4 Intérêts et limites des mesures faites en ville

L'ensemble de ces dispositifs permet d'aborder l'ESU par une nouvelle approche en s'affranchissant des limitation liées à la modélisation des sources et de leur propagation dans l'environnement urbain. Si les mesures participatives permettent d'estimer des niveaux sonores toutes sources confondues, les réseaux de capteurs et les mesures mobiles permettent une meilleure description dynamique et spatiale des ESU qui sont impossibles à obtenir avec les modèles prédictifs. Leurs utilisations offrent donc une représentation globale des ESU et ouvrent donc la voie vers de nombreuses applications :

— estimation des niveaux sonores du trafic et amélioration de la cartographie de bruit,

— identification et détection des sources sonores spécifiques,

— évaluation et classification plus complète des ESU par des indicateurs physiques,

— et représentation possible des ESU selon la perception des citadins.

Ces méthodes ne sont toutefois pas exemptes de défauts. Les réseaux de capteurs sont des systèmes complexes à gérer par leur installation et leur entretien. De plus, la question de l'interpolation entre les points de mesures reste une source d'approximation. À l'inverse, les mesures mobiles permettent de mieux estimer les variations spatiales aux dépends des variations à long-terme. Mais elles restent très couteuses en temps à réaliser à l'échelle d'une ville. Enfin les mesures participatives présentent de nombreuses incertitudes quant à la qualité de la mesure dues aux performances des capteurs des smartphones ou de la mesure réalisée qui nécessitent un traitement du signal important.

Toutefois, s'il existe déjà des outils destinés à évaluer les ambiances sonores urbaines ou qui lient leur perception par les citadins à des indicateurs physiques, la description des ESU selon les différentes sources sonores présentes nécessitent de disposer d'outils de traitements du signal adaptés afin d'en extraire leur contributions. Or, l'ESU est un milieu complexe, composé d'une multitude de sources variées (trafic routier, voix, oiseaux, klaxon, bruit de pas...) dont leurs allures temporelles (parfois brèves pour le retentissement d'un klaxon ou longues pour le passage d'une voiture) et fréquentielles (dans les basses fréquences pour le trafic, dans les hautes

fréquences pour le sifflement des oiseaux) diffèrent, voir Figure 1.7. L'ensemble de ces sources est aussi susceptible d'être généré simultanément. La création d'outils adaptés à cet environnement n'est donc pas triviale.

Des outils d'identification ou de détection ont déjà été développés pour des sons environnementaux [Mesaros *et al.*, 2015, Chachada et Kuo, 2014, Cakir *et al.*, 2015], mais la tâche de séparation de tels signaux au sein de mélanges sonores urbaines reste, quant à elle, pour l'instant peu étudiée. Dans le cas d'étude perceptive, réussir à isoler et caractériser les différentes sources sonores seraient très utile afin de relier l'évaluation perceptive des citadins réalisé lors de marches sonores non plus à des niveaux sonores globaux mais soit à celui de certaines sources ou bien en fonction de leur temps de présence comme dans [Aumond *et al.*, 2017a]. Développer de tels outils serait également nécessaire et utile, pour l'amélioration de la cartographie du bruit de trafic par exemple. Car s'il existe des endroits où celui-ci est prépondérant sur les autres sources sonores (périphérique, grand boulevard) et donc que son niveau sonore peut être estimé facilement, de nombreux autres lieux (dans des rues calmes, au niveau de parc) contiennent majoritairement d'autres sources sonores (voix, oiseaux...). Ne pas réussir à isoler la contribution du trafic routier des autres sources dans ces environnements risque alors de mener à de mauvaises estimations de son niveau sonore et de son temps de présence.

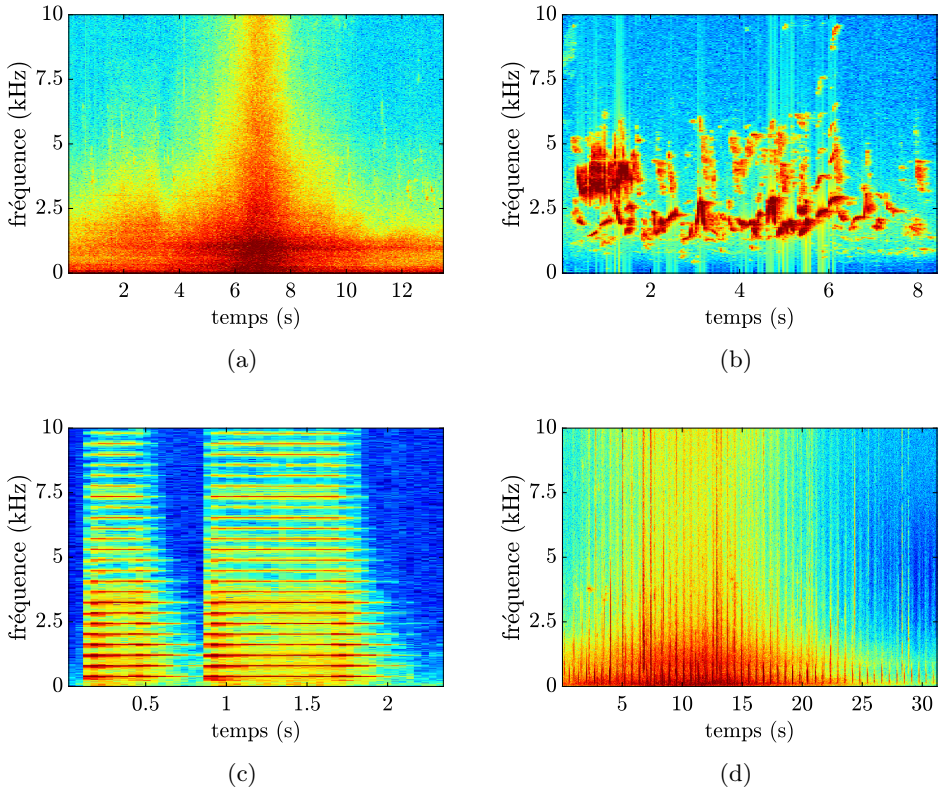


FIGURE 1.7 – Spectrogrammes d'un passage d'une voiture (a), d'un sifflement d'oiseaux (b), d'un klaxon (c) et d'un bruit de pas (d).

Il est donc nécessaire de générer un outil adapté à l'ESU afin de pouvoir extraire les contributions et les niveaux sonores des sources présentes en ville à partir de mesures et d'enregistrements. En conséquence, les travaux de cette thèse cherchent à répondre à ces questions :

— **Comment déterminer le niveau sonore du trafic routier en ville ? est-il possible de déterminer d'autres sources sonores ?**

— **Quelles sont les méthodes disponibles pour réaliser cette tâche ? Quel est la méthode la plus adaptée à notre environnement d'étude parmi celles existantes ?**

— **Quel protocole expérimental mettre en place pour tester et valider les performances de cet outil ?**

1.4 Estimation du niveau sonore du trafic routier

Étant la source principale de bruit en ville ainsi que la plus gênante [European Environment Agency, 2014], le trafic routier sera la source d'intérêt étudiée dans ce document. Le principe général de la méthode proposée est résumé en Figure 1.8 : à partir d'un enregistrement audio monophonique (réalisé en format wav et de fréquence d'échantillonnage 44,1 kHz), un outil, appelé *estimateur*, détermine le niveau sonore estimé du trafic routier. L'objectif est donc de construire cet estimateur et un protocole expérimental adéquat afin de générer l'erreur d'estimation la plus faible possible du niveau sonore du trafic.

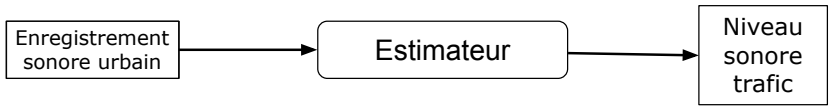


FIGURE 1.8 – Vue synthétique de l'approche proposée.

Parmi les travaux qui s'intéressent au trafic routier à partir d'enregistrements sonores d'ESU, des outils de reconnaissance du bruit [Defréville *et al.*, 2006], d'estimation du débit véhicule [Torija et Ruiz, 2012], de détection d'accidents de la route [Harlow et Wang, 2001] ou bien d'estimation du signal*trafic* basé sur l'antennerie de microphones [Leiba *et al.*, 2017] ont déjà été développés. La détermination du niveau sonore du trafic routier en tenant compte des autres sources sonores présentes parmi ces enregistrements a pour l'instant été très peu étudiée. On peut citer les travaux réalisés récemment au sein du projet DYNAMAP [Socoró *et al.*, 2017]. Leur approche consiste à entrainer une méthode de détection en vue d'estimer les trames temporelles où la classe de son *trafic* n'est pas présente afin de les rejeter lors de l'estimation des niveaux sonores. Une limite de cette approche est la possibilité de considérer des faux-positifs et de ne pas répondre à la question du recouvrement avec les autres sources sonores.

1.5 Méthode proposée

Ici, l'approche choisie est différente : l'estimateur du niveau sonore du trafic routier s'appuie sur une méthode de séparation de sources (voir chapitre 2) afin d'extraire l'intégralité de la composante *trafic* des enregistrements audio parmi les autres sources sonores présentes (voir Figure 1.9). L'un des intérêts est notamment la prise en compte naturelle du phénomène de recouvrement temporel. Cette méthode, en isolant la contribution *trafic* d'un mélange sonore permet de déterminer plusieurs de ses caractéristiques dont le niveau sonore. Cette estimation dépend de deux grandeurs : son unité et sa durée d'acquisition δ_t.

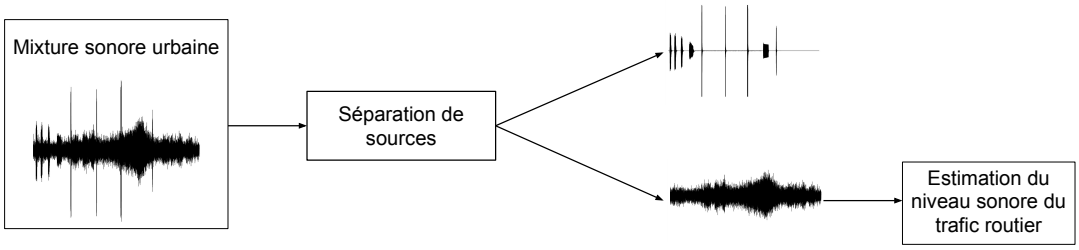


FIGURE 1.9 – Schéma de l'approche par séparation de sources d'un mélange sonore mélangeant une composante *trafic* avec des sons de klaxons.

Ce niveau sonore du trafic peut s'exprimer soit en grandeur linéaire (équivalent à la pression p_{rms} en Pascal), soit en grandeur logarithmique (en dB), tel que

$$L_p = 20 \times \log_{10}\left(\frac{p_{rms}}{p_0}\right) \tag{1.13}$$

avec $p_0 = 2 \times 10^{-5}$ Pa, la pression acoustique de référence. Pour ces travaux, le choix est fait d'exprimer les niveaux sonores en dB afin d'éviter de manipuler des valeurs dont les ordres de grandeurs seraient trop différents et de les calibrer. Enfin, le dB étant une grandeur couramment utilisée dans le domaine de l'acoustique urbaine, c'en est une représentation adaptée. La pondération A n'est pas considérée afin de se focaliser sur une estimation physique du niveau *trafic*. De plus, celle-ci diminue la présence des basses fréquences où se trouvent les composantes du trafic.

La durée d'intégration δ_t sert ensuite à déterminer un niveau sonore équivalent, L_{eq}, sur une plage de temps,

$$L_{eq} = 10 \times \log_{10}\left(\frac{1}{\delta_t}\int_t^{t+\delta_t} 10^{Lp/10} dt\right). \tag{1.14}$$

Quelle valeur choisir pour δ_t ? Choisit-on un niveau sonore exprimé toutes les 125 ms ? toutes les secondes ? toutes les minutes ? Aux vues des utilisations de cet estimateur qui sont envisagée (meilleur estimation de la contribution du trafic, amélioration des cartes de bruits en ville), le choix est fait de considérer une temps d'intégration long. Il n'est en effet pas pertinent de choisir un pas d'intégration trop faible (de 125 ms par exemple) qui serait alors une précision trop exi-

geante par rapport aux applications souhaitées. Pour le premier corpus (voir partie 4.4), la durée des scènes sonores est la même (30 secondes). La durée d'intégration δ_t sera alors définie selon leur longueur ($\delta_t = 30$ secondes). Puis, dans le second corpus (partie 4.5), la durée des scènes sonores étant variable, on choisira un temps d'intégration ramené à la minute ($\delta_t = 60$ secondes).

Il faut ensuite pouvoir comparer le niveau sonore estimé du trafic, $\tilde{L}_{eq,tr.,\delta_t}$, à sa valeur exacte, $L_{eq,tr.,\delta_t}$, afin de pouvoir évaluer la justesse de l'estimateur. En se basant sur des enregistrements sonores, dans le cas où il n'y a que du trafic, l'estimation fournie par la méthode peut être facilement comparée mais quid des scènes sonores où le trafic n'est pas prépondérant et est capté avec d'autres sources sonores ? La valeur exacte du trafic est l'inconnue qu'on cherche justement à déterminer. Sans cette référence, il est impossible de comparer la valeur estimée du trafic et ainsi la validité et les performances de l'estimateur. Le choix est donc fait d'utiliser non pas des enregistrements audio mais des scènes sonores issues d'un processus de simulation où un contrôle complet des classes sonores présentes, ainsi que de leur niveau sonore, est alors possible. Grâce à ce procédé, la valeur exacte, $L_{eq,tr.,\delta_t}$, est ainsi obtenue. La Figure 1.10 résume le schéma global du procédé suivi.

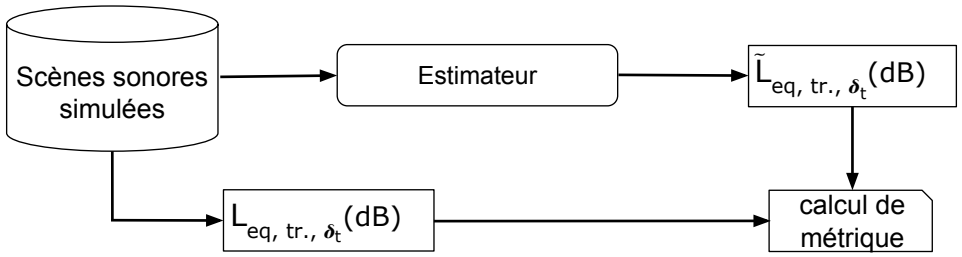


FIGURE 1.10 – Schéma-bloc du protocole expérimental.

Les niveaux sonores exacts et estimés du trafic routier sont ensuite comparés à travers un calcul de métrique. Le choix du calcul de métrique est un choix important puisqu'il conditionne les conclusions qui seront faites. Dans le cadre de la séparation de sources, plusieurs métriques ont été développées afin d'harmoniser et de faciliter la comparaison entre les performances des méthodes notamment dans [Vincent *et al.*, 2006]. Ici, puisque la performance de l'estimateur est liée à son erreur d'estimation du niveau sonore du trafic, c'est un indicateur d'écart qui est choisi. Parmi ces métriques qui composent cette classe (somme des carrés des résidus, la racine de l'erreur quadratique moyenne ($RMSE$ pour *Root Mean Square Error* en anglais), l'erreur absolue moyenne en pourcentage ($MAPE$)...), c'est l'erreur absolue moyenne, MAE (*Mean Absolute Error*) qui est retenue :

$$MAE = \frac{\sum_{i=1}^{M} |L^i_{eq,tr.,\delta t} - \tilde{L}^i_{eq,tr.,\delta t}|}{M}. \tag{1.15}$$

où M est le nombre de niveaux sonores calculés. Contrairement aux erreurs $RMSE$, qui revient à la racine carrée de la moyenne du carré des différences entre les données observées et réelles, ou $MAPE$, l'erreur moyenne en pourcentage, l'erreur MAE présente l'intérêt de

considérer un poids identique entre chaque différence et ainsi d'être moins sensible aux valeurs aberrantes. Par la réalisation de moyennes arithmétiques simples sur des grandeurs en dB, l'erreur MAE présente l'avantage de comparer équitablement les performances des estimateurs dans des cas variés. Par exemple, pour un estimateur qui estimerait systématiquement le niveau sonore du trafic moitié moins, que le niveau soit fort (94 dB) ou faible (54 dB), l'erreur, dans les deux cas, sera de 3 dB. Dans un cas linéaire, cette erreur s'exprime à partir des valeurs efficaces de la pression,

$$MAE(Pa) = \frac{\sum_{i=1}^{M} |p^i_{rms,trafic,\delta t} - \tilde{p}^i_{rms,trafic,\delta t}|}{M}. \tag{1.16}$$

Dans cet exemple, les erreurs seraient respectivement de 0,5 Pa et de 0,05 Pa. Le calcul d'erreur serait alors plus impacté par les erreurs réalisées dans les forts niveaux. Pour mieux équilibrer cette disparité, l'utilisation de cette métrique avec des grandeurs en dB est cohérente et adaptée à ces travaux. Enfin, si l'usage d'opérateurs statistiques sur des grandeurs logarithmiques peut être surprenant, en vue d'un traitement statistique des mesures, celui-ci est courant dans la communauté d'acoustique environnementale [Aumond *et al.*, 2018a, Morillas et Gajardo, 2014].

1.6 Conclusion du chapitre

Les environnements sonores urbains sont ainsi des milieux complexes qu'il est nécessaire de savoir caractériser afin de mieux connaitre les zones où les niveaux sont élevés et ceux où ils sont faibles. Les outils les plus utilisés sont, à l'heure actuelle, basés sur des modèles d'émissions et de propagations sonores qui permettent de prédire les niveaux sonores de sources spécifiques (bruit de trafic routier, ferroviaire et aérien et des ICPE). Ces modèles limitent toutefois la connaissance des ESU dans leur globalité en se focalisant uniquement sur des sources de bruit. De plus, ils présentent des limites liés à la réduction du nombre de catégories de voitures ou de revêtements, aux limites des modèles de propagation ou à la modélisation numérique. Ainsi le déploiement de réseaux de capteurs en ville et l'émergence des mesures participatives sont des pistes intéressantes en vue d'obtenir une caractérisation plus globale mais aussi plus fine de ces environnements. Ces approches s'affranchissent des difficultés techniques rencontrées lors des simulations et permettent de prendre en compte l'intégralité de l'ESU. Seulement pour estimer la contribution des sources sonores de l'ESU, il est nécessaire de disposer d'outils de traitement du signal adaptés à ces environnements et à ces sources. Encore peu d'études se sont focalisées sur cette question. C'est pourquoi les travaux de cette thèse s'attache à générer un outil basé sur la séparation de sources en vue d'estimer le niveau sonore du trafic routier, source de bruit principal en ville et la plus gênante auprès des citadins.

À partir de cette proposition, il reste maintenant à :

— **déterminer quelle méthode de séparation de sources choisir comme estimateur,**

— **construire correctement des corpus de scènes sonores urbaines pour tester le comportement et l'efficacité de l'estimateur retenu.**

Chapitre 2

Méthodes de séparation des sources sonores

Résumé

Les méthodes de séparation de sources sont des outils visant à extraire les différentes composantes sonores d'un enregistrement audio. Différentes approches existent et sont décrites pour ensuite être comparées. Parmi ces méthodes, la Factorisation en Matrices Nonnégatives se révèle être la plus adaptée pour mener les travaux de cette thèse en cela qu'elle intègre naturellement dans son fonctionnement le recouvrement temporel entre différentes sources sonores et est adaptée aux enregistrements audio monophoniques.

Savoir isoler les différentes composantes qui constituent un signal est une question complexe intervenant dans des domaines variés comme en biologie [Chiappetta *et al.*, 2004], dans le monde médical [Jung *et al.*, 2000] ou bien encore dans le domaine de l'image [Nuzillard et Bijaoui, 2000]. En audio, cette question intervient alors que la quasi-totalité des environnements sonores qui nous entourent sont composés d'une multitude de sources sonores. En considérant N sources sonores qui possèdent leurs propres caractéristiques acoustiques (intensité, fréquences, durée), reçu par un capteur $x_(i)$, le problème de la séparation de sources peut se poser comme un problème linéaire inverse. Le cas le plus simple est celui qui suppose des mixtures instantanées linéaire,

$$x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t), \tag{2.1}$$

où $x_i(t)$ est un signal capté à l'instant t au point i, $s_j(t)$ une source originale j émise à l'instant t et a_{ij}, la contribution de la source j sur le signal situé en i. On considère alors que les

signaux des N sources arrivent au même moment à chaque capteur. Une approche plus complexe et plus complète est de considérer le problème à partir de mixtures convolutives qui considèrent les multiples voies de propagation des sources s_j (champ direct et réfléchis) jusqu'au capteur x_i,

$$x_i(t) = \sum_{j=1}^{N} \sum_{k=1}^{+\infty} a_{ijk} s_j(t), \tag{2.2a}$$

$$= \sum_{j=1}^{N} a_{ij}(t) * s_j(t). \tag{2.2b}$$

avec $a_{ij}(t)$, la réponse impulsionnelle du filtre a_{ij} considéré comme un Système Linéaire Invariant. Chaque chemin de propagation est caractérisé un coeffcient a_{ijk} qui prend en compte l'atténuation énergétique et le temps de propagation entre la source s_j et le capteur x_i. Ces deux approches peuvent être résumées selon [Cardoso, 1998] par la relation

$$x_i(t) = \sum_{j=1}^{N} S_j(t) \tag{2.3}$$

qui décompose le signal capté i comme la contribution de chacune des N sources modulées par l'environnement. $S_j(t)$, appelé *image spatiale de la source* j. Notons que le problème émis par la tâche de la séparation de sources est équivalent à la définition, au début du chapitre 1, d'un ESU où le filtre mixant $a_{ij}(t)$ de l'équation 2.2b correspond au filtre de propagation $\delta(t)$ (équation1.1b).

L'intérêt de développer des outils de séparation de sources sur des enregistrements audio est de pouvoir éditer, analyser ou modifier les composantes extraites. Ceci est utile pour des applications comme, par exemple,

— le débruitage de la voix pour améliorer la qualité des appareils auditifs [Gannot *et al.*, 2017],

— la transcription des mélodies d'un morceau de musique [Vincent, 2006],

— la restauration d'enregistrements audio [Cañadas-Quesada *et al.*, 2016].

De par la variété des sons, c'est donc une tâche complexe qui est à l'étude depuis plus de 30 ans. De ces recherches, plusieurs méthodes ont émergé qui diffèrent selon les applications visées. On propose dans ce chapitre de présenter certaines des méthodes les plus couramment utilisées afin de déterminer celle qui est la plus adaptée à notre cas d'étude.

2.1 Analyse Computationnelle de Scènes Auditives

L'Analyse de Scènes Audio Computationnelle (abrégé CASA pour *Computational Auditory Scene Analysis* en anglais) est une des premières techniques numériques cherchant à séparer les différentes sources composant un signal. Elle fut proposée par Brown et Cooke [Brown et Cooke, 1994] et se base sur la simulation de la réponse auditive humaine. La méthode CASA est inspirée de l'Analyse de Scènes Auditives de Bregman [Bregman, 1990] qui explore les façons

dont le cerveau humain comprend et organise les sons qui l'entourent. L'architecture de la CASA se décompose en 4 parties [Wang et Brown, 2006] :

— un filtrage cochléaire qui consiste en une suite de filtres passe-bas qui modélisent l'oreille externe et moyenne, et d'un filtre gammatone qui simule les réponses impulsionnelles de chaque cellule ciliée. Le signal obtenu est exprimé, en sortie, au travers d'un cochléogramme.

— Une analyse temps-fréquence qui permet, au travers différents outils, d'augmenter les dimensions du problème et de mettre en évidence la présence de sons harmoniques notamment :

  - la corrélation croisée entre les canaux fréquentiels proches pour faire émerger la présence des formants,
  - la corrélation croisée entre les deux canaux des deux capteurs pour localiser la source grâce à leur déphasage,
  - la fonction d'autocorrélation dans chaque canal pour faire émerger des maximas à des positions correspondant aux périodes d'un son,
  - un lissage temporel afin de faire apparaitre des phénomènes de modulation.

— Un groupement de sources qui ré-organise ensuite les objets élémentaires pour construire les sources sonores en appliquant, par exemple, une contrainte temporelle sur les représentations spectrales. Ce groupement peut se faire à partir de stimuli (CASA de type *bottom-up*) ou bien à l'aide de schéma déjà établi (CASA de type *top-down*).

— Un masquage binaire temps-fréquence construit pour chaque source identifiée qui, appliqué sur le spectrogramme initial, permet d'isoler les différentes sources sonores.

Développée à partir de la compréhension de certains aspects des capacités d'analyse des sons par notre cerveau, la méthode CASA a notamment trouvé des applications dans le domaine de la parole [Ellis, 1999, Brown et Wang, 2005, Shao *et al.*, 2010] ou pour la reconnaissance de scènes sonores [Peltonen *et al.*, 2002].

2.2 Algorithme DUET

L'algorithme de séparation de sources DUET (*Degenerate Unmixing Estimation Techniques)* est une méthode proposée par [Rickard, 2007] qui permet de déterminer N sources sonores d'une mixture sonore à partir du déphasage et de l'atténuation entre les enregistrements de deux microphones $x_1(t)$ et $x_2(t)$. Cette approche se base sur 2 hypothèses : les signaux sont émis dans des conditions anéchoïques et il n'y a pas (ou très peu) de recouvrement fréquentiel entre les N sources sonores $s_j(t)$. La condition d'anéchoïcité permet d'exprimer les 2 signaux captés sous la forme d'un signal exprimé en champ direct, $x_1(t) = \sum_{j=1}^{N} s_j(t)$, et d'un autre déphasé, $x_2(t) = \sum_{j=1}^{N} a_j s_j(t - \delta_j)$, pondéré par l'atténuation a_j et en déphasage (positif ou négatif) δ_j

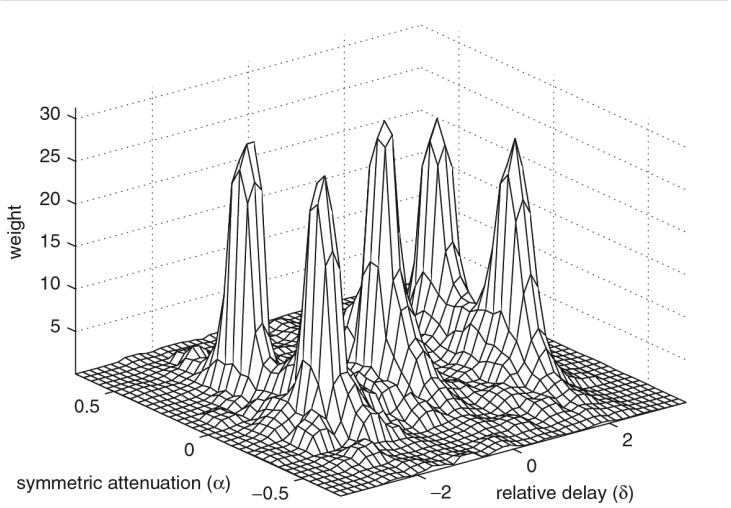


FIGURE 2.1 – Histogramme 2D pour un signal sonore composé de 5 sources sonores, générant 5 pics [Rickard, 2007].

avec la condition $\delta_j \leqslant \frac{d_{\mu_1,\mu_2}}{c_0}$ où d_{μ_1,μ_2} est la distance entre les 2 microphones et c_0 la célérité du son. En raison de la distance entre les microphones et du déphasage entre les signaux, le dispositif limite la fréquence maximale du signal traitée à $f_{max} = \frac{c_0}{2d_{\mu_1,\mu_2}}$. Le problème s'exprime sous forme matricielle :

$$\begin{bmatrix} X_1(\omega,\tau) \\ X_2(\omega,\tau) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-j\omega\delta_1} & \dots & a_N e^{-j\omega\delta_N} \end{bmatrix} \times \begin{bmatrix} S_1(\omega,\tau) \\ \vdots \\ S_N(\omega,\tau) \end{bmatrix} \tag{2.4}$$

avec X_i et S_j, les représentations temps-fréquences du signal x_i et de la source s_j, à l'instant τ et à la fréquence $2\pi f = \omega$. Le rapport des amplitudes des signaux permet d'exprimer les rapports d'amplitudes $\tilde{a}$ et de phases $\tilde{\delta}$:

$$\tilde{a}(\omega,\tau) = \left|{}^{X_2(\omega,\tau)}/_{X_1(\omega,\tau)}\right|, \tag{2.5a}$$

$$\tilde{\delta}(\omega,\tau) = {}^{-1}/_{\omega}\underline{\angle\left({}^{X_2(\omega,\tau)}/_{X_1(\omega,\tau)}\right)}. \tag{2.5b}$$

avec $\underline{\angle {}^{a}/_{b}}$, la phase du rapport des nombres complexes a et b. L'ensemble des valeurs obtenues est ensuite exprimé au sein d'un histogramme à 2 dimensions ($\tilde{a}$ et $\tilde{\delta}$) (voir Figure 2.1). Les différents pics émergeant associés à un couple particulier de $\tilde{a}$ et $\tilde{\delta}$ permettent alors de générer un masque binaire en temps-fréquence afin d'extraire la source sonore qui y est associée.

Une implémentation de cet algorithme, pour une utilisation en temps réel, est proposée dans [Rickard *et al.*, 2001]. Afin de s'extraire des conditions strictes imposées par la méthode et de s'approcher d'applications plus réelles, [Rickard, 2007] propose de considérer l'ajout d'un terme de bruit à l'équation 2.4 et de déterminer l'*atténuation symétrique* α_j et le *déphasage relatif* δ_j de chaque source :

$$\alpha_j = \frac{\iint_{(\omega,\tau)} |X_1(\omega,\tau)X_2(\omega,\tau)|^p \omega^q \alpha(\omega,\tau) d\omega d\tau}{\iint_{(\omega,\tau)} |X_1(\omega,\tau)X_2(\omega,\tau)|^p \omega^q d\omega d\tau}, \tag{2.6}$$

avec $\alpha(\omega,\tau) = \left|\frac{X_2(\omega,\tau)}{X_1(\omega,\tau)}\right| - \left|\frac{X_1(\omega,\tau)}{X_2(\omega,\tau)}\right|$ et

$$\delta_j = \frac{\iint_{(\omega,\tau)} |X_1(\omega,\tau)X_2(\omega,\tau)|^p \omega^q \tilde{\delta}(\omega,\tau) d\omega d\tau}{\iint_{(\omega,\tau)} |X_1(\omega,\tau)X_2(\omega,\tau)|^p \omega^q d\omega d\tau}. \tag{2.7}$$

Ces deux indicateurs dépendent de deux constantes définies, p et q, qui permettent de pondérer le poids de chaque point temps-fréquence :

— $p = 0$, $q = 0$ revient à l'algorithme DUET original,

— $p = 1$, $q = 0$ donne plus de poids à la reconstruction de l'atténuation symétrique α_j [Yilmaz et Rickard, 2004],

— $p = 1$, $q = 2$ donne plus de poids à la reconstruction du déphasage relatif δ_j [Yilmaz et Rickard, 2004],

— $p = 2$, $q = 2$ est adapté aux signaux ayant un faible rapport signal à bruit ou bien pour des mixtures de paroles [Melia et Rickard, 2007].

Différentes applications basées sur cet algorithme existent notamment pour des signaux de paroles [Yilmaz et Rickard, 2004, Jourjine *et al.*, 2000]. Plusieurs développements ont été proposés afin d'étendre l'utilisation de cette méthode. Dans [Melia et Rickard, 2007] plusieurs versions sont décrites et comparées comme la *echoic DESPRIT* qui propose avec M microphones de retrouver N signaux (en supposant que le nombre de chemins empruntés par les signaux est inférieur à $M/2$) ou bien la *echoic ESPRIT*, où il est supposé que le nombre de sources actives simultanément n'excède pas $M/2$.

2.3 Analyse en Composantes Indépendantes

L'Analyse en Composantes Indépendantes (ACI) [Comon, 1994, Jutten et Herault, 1991] est une méthode appartenant aux méthodes dites de *séparation de sources aveugle*, c'est-à-dire qui séparent un ensemble de sources sonores d'une mixture sans (ou avec peu) informations sur celles-ci. L'illustration la plus couramment citée, pour cette méthode, est l'effet « cocktail party ». Cet effet résume le processus qui permet à un être humain de séparer la voix de l'interlocuteur, avec qui il discute, du flux sonore environnant composé d'autre discussions, de musique... Cette capacité est notamment permise par l'indépendance entre le signal *voix* et les autres sources sonores aux alentours ainsi que par l'écoute binaurale du sujet. L'ACI se base sur ces hypothèses et s'exprime alors sous la forme d'un produit matriciel :

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.8}$$

où $\mathbf{x}$, de dimensions $N \times 1$, exprime l'ensemble des mesures faites par N capteurs, $\mathbf{s}$, de dimensions $N \times 1$, résume les différentes sources présentes et $\mathbf{A}$, de dimensions $N \times N$, une matrice déterministe résumant les aspects de propagation entre les sources et les capteurs. $\mathbf{A}$ et $\mathbf{s}$ étant inconnus, plusieurs hypothèses sont considérées afin de résoudre le problème : i) $\mathbf{s}$ est composé de sources sonores indépendantes, ii) les sources sonores ne suivent pas de distribution gaussienne (ou pas plus de une), iii), $\mathbf{A}$ est une matrice carrée inversible. Le problème s'exprime alors sous la forme :

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}. \tag{2.9}$$

L'ACI peut être vue comme une extension de l'Analyse en Composante Principale (ACP) où il y est supposé, non pas l'indépendance des composantes, mais leur décorrélation et où sont déterminées les grandeurs qui ont le plus de variances. De nombreux développements ont été proposés afin d'obtenir une indépendance maximale entre les sources :

— la minimisation de l'information mutuelle [Hyvärinen, 1997] qui consiste à maximiser un calcul de distance (au travers d'une divergence de Kullback-Leibler) entre l'entropie de l'ensemble des composantes et celle d'une composante i.

— La décorrélation non-linéaire qui consiste à minimiser la corrélation entre deux composantes y_i et y_j chacune exprimée par une fonction $f(y)$ et $g(y)$ dont au moins une est non-linaire (polynôme de degré 2 ou plus, fonction tangente hyperbolique...). Pour passer de la non-corrélation à l'indépendance, [Jutten et Herault, 1991] ajoute une condition où la fonction non-linéaire (ou une des deux) est impaire avec une moyenne nulle.

— lala maximisation de la « non gaussianité » des composantes, basée sur le *Théorème Centrale Limite* qui démontre que la somme de variables aléatoires indépendantes définies dans un même espace tend vers une distribution gaussienne. La recherche de l'indépendance des sources revient alors à maximiser la « non-gaussianité » entre les composantes. Cette approche est la base de la mesure de la négentropie (ou entropie relative) [Lee *et al.*, 2000] ou de l'algorithme FastICA [Hyvärinen, 1999] qui est l'algorithme le plus couramment utilisé pour résoudre l'ACI.

Une revue exhaustive des méthodes employées est effectuée dans [Hyvärinen *et al.*, 2004]. Pour être résolue, l'ACI est, le plus souvent, sur-déterminée, ce qui signifie qu'il y a au moins autant ou plus de signaux captés que de sources sonores. Pour obtenir une matrice $\mathbf{A}$ carrée, une ACP peut être utilisée afin de supprimer les informations redondantes. Le cas de la sous-détermination a toutefois été étudié, par exemple dans [Bofill et Zibulevsky, 2000] en considérant les signaux comme parcimonieux. L'ACI a trouvé de nombreuses applications, en tant que méthode de séparation de sources aveugles, dans le domaine médical, pour extraire les différentes composantes d'un signal d'électroencéphalogramme [Delorme *et al.*, 2007, Makeig *et al.*, 1996] ou d'une IMR [Lee *et al.*, 1999], pour traiter des signaux contenant de la parole [Särelä et Valpola, 2005, Hsieh *et al.*, 2009] ou pour des contenus musicaux [Uhle *et al.*, 2003, Abdallah et Plumbley, 2003]. Des utilisations de l'ACI pour des sons environnementaux existent aussi [Lombard

et al., 2011, Eronen *et al.*, 2006]. Cette méthode peut également être utilisée pour des antennes acoustiques et pour la formation de voies (*beamforming*)[Cardoso, 1998, Saruwatari *et al.*, 2003].

La détermination des matrices **A** et **s** n'est toutefois pas sans générer des ambiguïtés bien identifiées dans la littérature. La première est l'*ambiguïté de permutation* qui traduit la variabilité dans l'ordre de détermination des composantes indépendantes : une composante estimée s_1 peut être déterminée à un rang différent sans pour autant changer la reconstruction du signal global. Ce problème est toutefois sans conséquence la plupart du temps. Une seconde limite est l'*ambiguïté d'échelle* qui traduit la possibilité d'avoir un facteur d'échelle présent dans la matrice **A** et **s** tel que :

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \alpha_1 a_{11} & \dots & \alpha_1 a_{1N} \\ \vdots & \ddots & \vdots \\ \alpha_N a_{N1} & \dots & \alpha_N a_{NN} \end{bmatrix} \times \begin{bmatrix} s_1/\alpha_1 \\ \vdots \\ s_N/\alpha_N \end{bmatrix}. \tag{2.10}$$

La détermination de l'amplitude exacte des signaux est donc soumise à l'incertitude [Naik et Kumar, 2011] mais peut être toutefois soulevée en normalisant les énergies des sources.

2.4 Factorisation en Matrices Non-négatives

La Factorisation en Matrices Non-négatives (abrégée NMF pour *Non-negative Matrix Factorization* en anglais) [Lee et Seung, 1999], appliquée à l'analyse d'un signal audio-numérique, est une méthode qui est basée sur une représentation linéaire de données :

$$\mathbf{V} \approx \mathbf{WH} \tag{2.11}$$

où **V** est le spectrogramme en amplitude ou en puissance d'un signal audio de dimensions $F \times N$, **W** est appelé *dictionnaire* ou *base*. De dimensions $F \times K$, il contient un ensemble de spectres sonores. **H**, de dimensions $K \times N$ est la matrice d'activation qui traduit les variations dans le temps de chaque spectre de **W**. K définit le rang des matrices et est le plus souvent choisi tel que $F \times K + K \times N \ll F \times N$ afin d'être une méthode qui permet la réduction de données. Là où l'ACI impose une contrainte d'indépendance, la NMF impose celle de la « non-négativité » (**V**, **W** et **H** $\in \mathbb{R}_+$). Cette contrainte n'autorise alors que des combinaisons additives entre les composantes et leur assure ainsi d'appartenir au même domaine et donc de leur interprétabilité. L'approximation entre **V** et **WH** est résolue en minimisant leur β-divergence :

$$\min D_\beta(\mathbf{V}||\mathbf{WH}) \quad \text{avec} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0. \tag{2.12}$$

Plusieurs approches existent afin de résoudre le problème 2.12, basées sur des approches itératives : mises à jour multiplicatives [Lee et Seung, 2000], méthode des moindres carrés alternés [Cichocki et Zdunek, 2007], gradient projeté [Lin, 2007]... La NMF permet aussi d'ajouter des contraintes sur chaque matrice [Virtanen, 2007, Bertin *et al.*, 2010] afin de forcer un type de comportement particulier des matrices suivant le problème posé.

Cette méthode a reçu une grande popularité en trouvant de nombreuses applications pour la musique (transcription de partitions de musique [Smaragdis et Brown, 2003, Bertin *et al.*, 2009], classification de genre musicaux [Panagakis *et al.*, 2008] et d'instruments [Benetos *et al.*, 2006], séparation des pistes musicales [Souviraà-Labastie *et al.*, 2015]) et pour la parole (débruitage [Wilson *et al.*, 2008a, Sprechmann *et al.*, 2014], séparation de sources [Smaragdis, 2007, Hurmalainen *et al.*, 2012])... Cette méthode a également déjà été confrontée à des sons environnementaux pour différentes tâches comme dans [Kumar *et al.*, 2016] où la NMF est utilisée afin de localiser l'origine de différents extraits sonores. Dans [Sobieraj *et al.*, 2017], la NMF sert à la détection d'oiseaux dans différents environnements sonores. Dans [Heittola *et al.*, 2011], la NMF est utilisée en vue de détecter plusieurs évènements sonores émis simultanément en représentant les évènements à travers des Chaines de Markov Cachées (HMM) et des Mel Frequency Cepstral Coeffcient (MFCC). Enfin dans [Innami et Kasai, 2012], la NMF est utilisée en tant que méthode de séparation de sources pour extraire différents signaux (voix, aboiement de chien, croassements de grenouille). Par son fonctionnement, la NMF présente l'avantage, en plus de prendre naturellement en compte le recouvrement temporel entre les sources sonores, d'être adaptée à des réseaux de capteurs monophoniques.

2.5 Détection d'évènements sonores

Cette approche n'est pas une méthode de séparation de sources puisque l'objectif est d'identifier les sources sonores qui composent une mixture pour ensuite ne considérer que celles d'intérêt.

La reconnaissance et la détection de sources sonores environnementales à l'aide de descripteurs [Dufaux *et al.*, 2000, Defréville *et al.*, 2006] reçoit une attention croissante. Ces travaux suivent le plus souvent un protocole établi : i) apprentissage d'un détecteur sur une base de données, ii) application de ce détecteur sur une base de test en vue d'estimer les performances de l'outil. Dans [Mesaros *et al.*, 2010], 61 évènements sonores sont détectés parmi des enregistrements audio annotés à l'aide de l'algorithme de Viterbi et des HMM. [Ntalampiras *et al.*, 2011] détecte des évènements anormaux tel que des cris, des tirs d'armes à feu à partir de mixtures sonores synthétisées où le niveau sonore relatif de ces évènements par rapport aux bruits de fond de la scène est échelonné entre -5 dB et 15 dB. Les descripteurs utilisés sont les MFCC, MPEG-7 et les *Perceptual Wavelet Packets* et le classifieur est basé sur un Modèle de Mixtures de Gaussiennes (GMM) et sur des HMM. Ces travaux seront étendus dans [Ntalampiras, 2014] pour réaliser un outil de surveillance du trafic routier. Dans le cadre du projet DYNAMAP, un détecteur d'évènements sonores anormaux (DESA) a été réalisé [Socoró *et al.*, 2017]. Celui-ci consiste, pour chaque trame temporelle, à y identifier la classe de son principale. Celles qui ne sont pas reconnues comme *trafic* sont alors rejetées et non prises en compte dans l'estimation du niveau sonore du trafic routier. Cet outil se base, là aussi, sur les MFCC (13 bandes), exprimés sur des trames temporelles de 30 ms, couplés à des GMM. Le corpus de test est divisé en deux classes : *urbain* et *banlieue*. Les performances du détecteur, évalué au travers du F-score, est pour chaque classe du corpus de 61,5 % et de 72,3 %.

2.6 Comparaison des approches

Le Tableau 2.1 compare les différentes caractéristiques requises par l'outil de séparation de sources afin de satisfaire le cahier des charges proposé dans la partie 1.4. Pour rappel, celui-ci doit être adapté aux sons environnementaux présentant du recouvrement temporel et aux réseaux de capteurs monophoniques.

Tableau 2.1 – Comparaison des 4 méthodes de séparation de sources et de la méthode DESA en vue d'estimer le niveau sonore du trafic routier à partir d'enregistrements monophoniques (« - » : pas adapté, « + » : adapté, « ++ » : très adapté).

	sons environnementaux	recouvrement temporel	réseau de capteurs monophoniques
CASA	-	+	-
Algorithme DUET	-	+	-
ACI	+	++	-
NMF	++	++	++
DESA	++	-	++

La méthode CASA et l'algorithme DUET sont des méthodes qui sont peu adaptées au cas d'étude présent, car elles sont développées notamment pour des sons harmoniques et qui nécessitent au minimum deux microphones pour chaque point de mesure ce qui n'est pas compatible avec des réseaux de capteurs monophoniques. La méthode CASA a toutefois été adaptée pour des enregistrements monophoniques [Hu et Wang, 2006] mais reste majoritairement associée aux sons stéréophoniques. De plus, basée sur une hypothèse d'anéchoïcité, malgré des extensions pour s'affranchir de celle-ci, la méthode DUET reste peu évidente et adaptée pour un environnement sonore urbain où le champs diffus est prépondérant. Si l'ACI est utilisé pour des mixtures sonores urbaines, elle l'est pour des antennes de microphones et la formation de voies. Là encore, cette approche n'est donc pas une méthode adaptée à la mise en place de réseaux de capteurs monophoniques. De plus l'*ambiguïté d'échelle* est un frein pour estimer correctement le niveau sonore du trafic. L'outil DESA répond à une grande partie des besoins, mais génère les problèmes classiques de détection de faux-positifs (des trames contenant des composantes *trafic* mais qui sont identifiées comme *interférante*). De plus, les cas où le trafic et une autre source *interférante* sont présents dans une même trame posent la question de son possible rejet. Ne pas considérer un ensemble de trames où le trafic est pourtant présent (mais rejeté parce qu'elle n'est pas la source principale) peut mener à des estimations du niveau sonore du trafic qui serait sous-estimées. Dans le cas où le trafic est la source sonore principale, aux abords des boulevards ou d'un périphérique, ce problème est limité. Mais dans les cas où les rues sont calmes, dans les parcs ou les places où les voix sont plus importantes, même si le trafic est moins présent, sa contribution à l'ESU reste présente et doit être estimée. Ainsi, c'est donc, la NMF qui, par son fonctionnement (reconstruire le spectrogramme d'un signal audio à l'aide d'un dictionnaire), paraît être la méthode la plus adaptée à ces travaux. Celle-ci se base sur un enregistrement audio réalisé par un seul capteur, donc compatible avec les réseaux de capteurs monophoniques. Le

recouvrement est naturellement pris en compte par la contrainte de non-négativité qui n'autorise que des combinaisons additives. Ainsi, l'ensemble des sources présentes dans **W** peuvent être, pour une trame donnée, toutes considérées simultanément. Enfin, son application à des sons environnementaux dans de précédentes études, même si le trafic n'en est pas la source sonore principale, permet d'assurer la compatibilité de cette méthode face à des mixtures sonores urbaines.

Chapitre 3

La Factorisation en Matrices Non-négatives

Résumé

Le fonctionnement de la Factorisation en Matrices Non-négatives (NMF) est présenté dans ce chapitre. Cette méthode consiste à approximer le spectrogramme en amplitude d'un signal audio par le produit de deux matrices positives : W, un dictionnaire composé de spectres audio, et H, la matrice d'activation temporelle. Les différents aspects de son fonctionnement sont décrits : familles de divergences, algorithmes de mise à jour des matrices, méthode d'apprentissage du dictionnaire. Une forme de NMF est également proposée : la NMF *initialisée seuillée* où un dictionnaire, appris sur la source d'intérêt, est mis à jour et dont les éléments relatifs à cette source sont ensuite sélectionnés par une technique de seuillage. Enfin, les différentes contraintes, qui peuvent être apposées sur les matrices de la NMF, sont décrites et notamment la contrainte de régularité temporelle et de parcimonie.

La Factorisation en Matrices Non-négatives étant la méthode retenue pour la suite des travaux, ce chapitre en présente son fonctionnement. Les principes généraux sont dans un premier temps détaillés, puis les différentes approches de cette méthode sont explicitées.

3.1 Principe de fonctionnement de la Factorisation en Matrice Non-négatives

La Factorisation en Matrices Non-négatives (abrégée NMF pour *Non-negative Matrix Factorization* en anglais [1]) est une technique d'approximation linéaire visant à décomposer une

1. En anglais, le terme *non-negative* fait référence à des valeurs positive ou nulle. En français, le terme correspondant serait *positif*, la traduction littérale serait alors *Factorisation en Matrices à valeurs positives*. Nous

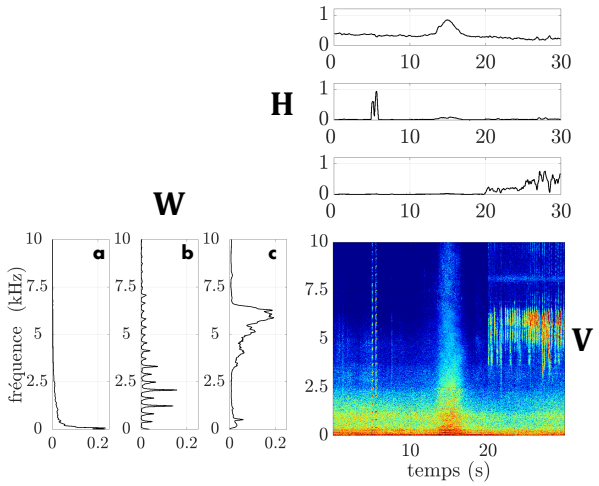


FIGURE 3.1 – Exemple d'une NMF pour un signal audio de mixture urbaine composé de 3 sources sonores. **W** et **H** sont constitués de 3 bases (K = 3) : a) un spectre voiture, b) un spectre de klaxon, c) un spectre d'oiseau.

matrice **V** non-négative de dimensions $F \times N$ en un produit de deux matrices tel que

$$\mathbf{V} \approx \mathbf{WH} \tag{3.1}$$

où **W** et **H** sont deux matrices, également non-négatives, de dimensions respectives $F \times K$ et $K \times N$, appelées *dictionnaire* et *matrice d'activation*. Le choix du rang de factorisation K est le plus souvent déterminé afin que la relation $F \times K + K \times N << F \times N$ soit respectée. Dans ce cas, la NMF est une méthode d'approximation dite de faible rang car elle permet la réduction de la dimensionalité des données.

La contrainte de non-négativité permet d'assurer seulement des combinaisons additives, la soustraction d'information n'est pas possible dans le produit **WH**. Cela assure ainsi que les éléments du dictionnaire appartiennent bien tous au même domaine non-négatif que les données d'observations **V**, et donc de leur interprétabilité. En cela, la NMF réalise alors une représentation dite « par partie » dans laquelle **W** est composée de briques élémentaires qui, additionnées, permettent d'approximer l'ensemble de **V**. Un exemple de NMF est présenté dans la Figure 3.1. Chaque colonne n de la matrice **V** est décrite comme la somme des éléments de **W** pondérés par la colonne n de la matrice **H** :

$$\mathbf{v} \approx \tilde{\mathbf{v}} = \mathbf{Wh} \tag{3.2}$$

où les caractères minuscules représentent des vecteurs colonnes et n est une trame temporelle. Les majuscules dénotent des matrices.

Cette méthode est assimilable aux méthodes de factorisation comme l'Analyse en Compo-

choisissons toutefois de considérer la traduction littérale en cela qu'elle est la plus couramment utilisée dans la communauté francophone du traitement du signal.

santes Principales (ACP), où la contrainte de non-négativité est remplacée par une contrainte d'orthogonalité entre les matrice **W** et **H**, ou l'Analyse en Composantes Indépendantes (ACI) où l'indépendance entre chaque composante est supposée et où il est possible d'obtenir des valeurs négatives.

Si la NMF fut introduite pour la première fois par Paatero et Tapper [Paatero et Tapper, 1994] en 1994 (mais sous le nom de *Positive Matrix Factorization*), elle doit sa popularité aux travaux de Lee et Seung [Lee et Seung, 1999] dont les résultats furent publiés dans la revue *Nature* en 1999. La NMF a ensuite trouvé de nombreuses applications dans des domaines variés : imagerie [Guillamet *et al.*, 2003, Monga et Mhcak, 2007], traitement de texte [Xu *et al.*, 2003, Berry et Browne, 2005], biologie [Gao et Church, 2005, Chen *et al.*,], gastronomie [Hawkins, 2006] ou bien encore pour la recommandation de contenus audio-visuels [Luo *et al.*, 2014]. Dans le domaine de l'audio, c'est Smaragdis et Brown [Smaragdis et Brown, 2003] qui furent les premiers à l'utiliser dans le cas de la transcription d'un morceau de musique polyphonique. Plus généralement, pour un signal audio quelconque, la NMF consiste à approximer son spectrogramme en amplitude ou en puissance, obtenu, par exemple, à partir d'une Transformée de Fourier à Court Terme, à l'aide du dictionnaire **W** constitué d'un ensemble de spectres sonores dont leurs amplitudes sont pondérées temporellement par les activateurs **H**.

De nombreuses applications de la NMF furent trouvées pour des signaux musicaux et contenant de la parole pour les tâches de détection [Dessein *et al.*, 2013], de reconnaissance de sources [Gemmeke *et al.*, 2013], de classification [Benetos *et al.*, 2006], de débruitage [Wilson *et al.*, 2008b] et de séparation de sources sonores [Virtanen, 2007]. Dans le cas de sons environnementaux (c'est-à-dire les sons qui ne sons ni musicaux ni de paroles), plusieurs applications ont été faites de la NMF. [Gemmeke *et al.*, 2013] utilisent la NMF en vue de réaliser de la détection d'évènements sonores sur des sons d'intérieurs (bruit de clés, sonnerie de téléphone, bruit de clavier...) et est testée sur une base de données synthétique où les évènements sonores sont artificiellement mixés à un bruit de fond, à des niveaux sonores calibrés. [Mesaros *et al.*, 2015] réalise également de la détection de sons environnementaux. Un aspect intéressant de leur étude est la comparaison des performances de leur outil à partir de la forme du dictionnaire selon les techniques de réduction de dimensions employées. Ils comparent le cas d'un dictionnaire composé de bandes fines mais réduit en nombre d'éléments K par un algorithme de clustering K-means et le cas d'un dictionnaire avec la base complète mais avec une représentation en bandes mel. Les auteurs observent alors que les deux techniques de réduction, sur leur corpus, offrent des performances similaires par rapport au dictionnaire original non réduit et suggèrent même la possibilité d'appliquer ces deux techniques de réduction en vue de réduire les tailles des matrices et ainsi réduire les coûts de calcul. Cet aspect sera repris pour la construction du dictionnaire dans la partie 5.3.1. [Innami et Kasai, 2012] ont proposé d'utiliser la NMF en vue de réaliser de la séparation de sources sur des mixtures sonores environnementales issues d'un processus de simulation. Leur méthode consiste dans un premier temps à isoler le bruit de fond des évènements sonores afin ensuite de les séparer individuellement. Pour cela, en plus d'une description

des éléments par des MFCC, deux autres paramètres descripteurs sont ajoutés : la variance de l'élément i sur la durée de la scène et la quantité d'éléments proches de zéro. Ces paramètres permettent de différencier les évènements sonores, dont ces valeurs seront alors élevées, du bruit de fond, où la variance et le rapport seront proches de zéro. Ces travaux sont toutefois, réalisés sur une base de données restreintes à quelques éléments audio.

3.2 Fonction de coût et familles de divergences

Le problème à résoudre lors d'une NMF est celui d'une minimisation où il faut trouver la combinaison optimale de **WH** qui sera la plus proche de **V**. Cela se traduit mathématiquement par la relation 3.3 :

$$\min D\left(\mathbf{V}\|\mathbf{WH}\right) \quad \text{avec} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0. \tag{3.3}$$

$D\left(\mathbf{V}||\mathbf{WH}\right)$ est alors une mesure de dissimilarité, appelée *fonction de coût*, qui peut appartenir à différentes familles de divergences comme les divergences de Csiszar [Cichocki *et al.*, 2006b] et de Bregman [Bregman, 1967, Dhillon et Sra, 2005]. Cette dernière famille de divergences est la plus couramment utilisée dans le cadre de la NMF. Elle se définit, dans un sous-ensemble convexe S d'un espace de Hilbert, comme :

$$D_{\Phi}(\mathbf{x}||\mathbf{y}) = \mathbf{\Phi}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\mathbf{\Phi}(\mathbf{y})\rangle \tag{3.4}$$

où $\mathbf{x} = (x_1, x_2, \ldots x_N)$ et $\mathbf{y} = (y_1, y_2, \ldots y_N)$ sont deux distributions, $\mathbf{\Phi}$ est une fonction continue dérivable et strictement convexe défini sur $\mathbb{R}^+$, $\nabla\mathbf{\Phi}(\mathbf{y})$ est le gradient de $\mathbf{\Phi}$ en $\mathbf{y}$ et $\langle ., . \rangle$ est le produit scalaire hermitien. L'équation (3.4) peut être décomposable élément par élément :

$$D_{\Phi}(\mathbf{x}||\mathbf{y}) = \sum_{n=1}^{N} d_{\Phi}(x_n|y_n) \tag{3.5}$$

avec $d_{\Phi}(x_n|y_n)$, la mesure de similarité entre deux scalaires et $\mathbf{\Phi}(\mathbf{x}) = \sum_{n=1}^{N} \phi(x_n)$. L'équation 3.4 se résume alors à une somme des divergences entre les composantes des distributions $\mathbf{x}$ et $\mathbf{y}$:

$$D_{\Phi}(\mathbf{x}||\mathbf{y}) = \sum_{n=1}^{N} \phi(x_n) - \phi(y_n) - \phi'(y_n)(x_n - y_n). \tag{3.6}$$

La fonction de coût s'exprime donc comme la divergence de Bregman appliquée à chaque élément de **V** et **WH**,

$$D\left(\mathbf{V}||\mathbf{WH}\right) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_{\phi}\left(\mathbf{V}_{fn}|\left[\mathbf{WH}\right]_{fn}\right). \tag{3.7}$$

Cette divergence possède plusieurs propriétés :

1. **Non-négativité** : $d_{\phi}(x|y) \geq 0$.

2. **Séparabilité** : si $d_\phi(x|y) = 0$ alors $x = y$.

3. **Convexité** : $D_\phi(\mathbf{x}||\mathbf{y})$ est une fonction réelle strictement convexe pour le 1er argument mais pas nécessairement pour le second.

4. **Linéarité** : pour deux fonctions convexes, réelles ϕ_1 et ϕ_2, $d_{\alpha\phi_1+\beta\phi_2}(x|y) = \alpha d_{\phi_1}(x|y) + \beta d_{\phi_2}(x|y)$ où α et $\beta \in \mathbb{R}^+$.

5. **Dualité** : pour la fonction convexe conjuguée ϕ^*, $d_{\phi^*}(y^*|x^*) = d_\phi(y|x)$.

3.3 Une sous-classe des divergences de Bregman : la β-divergence

Dans [Banerjee *et al.*, 2005], les auteurs démontrent qu'à chaque divergence de Bregman est associée une famille exponentielle unique $p\,(x|\theta, \lambda)$:

$$p\,(x|\theta,\lambda) = h(x,\lambda)\exp\left[\lambda^{-1}\left(\theta(y)x - \psi(\theta)\right)\right], \tag{3.8}$$

$$= g(x,\lambda)\exp\left(-\lambda^{-1}d_\phi(x|y)\right), \tag{3.9}$$

avec $h(x,\theta)$, la fonction de base, $\theta(y)$, le paramètre normal (ou canonique), λ, celui de la dispersion, $\psi(\theta)$ celui de la normalisation et $g(x,\lambda) = h(x,\lambda)\exp(\lambda^{-1}\phi(x))$. Les paramètres sont reliés entre eux par plusieurs relations $y(\theta) = \frac{d\varphi(\theta)}{d\theta}$, $\theta(y) = \frac{d\phi(y)}{dy}$ et $\frac{d\theta(y)}{dy} = v(y)^{-1}$. Un cas remarquable de distribution est la distribution de Tweedie [Jorgensen, 1987] qui relie la variance $v(x)$ à la moyenne (ou espérance) de la distribtion x par une relation polynomiale [Yilmaz et Cemgil, 2012] définie par un paramètre de forme β,

$$v(x) = x^{2-\beta}. \tag{3.10}$$

L'ensemble des divergences de Bregman définies par cette distribution dans l'équation 3.9 est alors paramétré par le choix de cette valeur β et peut être généralisé :

$$d_{\phi_\beta}(x|y) = \begin{cases} \dfrac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}), & \beta \in \mathbb{R}\backslash\{0,1\}, \quad (3.11a) \\ \dfrac{x}{y} - \log\dfrac{x}{y} - 1, & \beta = 0, \quad (3.11b) \\ x\log\dfrac{x}{y} - x + y, & \beta = 1. \quad (3.11c) \end{cases}$$

Les équations 3.11b et 3.11c sont des cas limites de l'expression 3.11a. Chaque paramètre de forme choisi coïncide alors avec une distribution particulière des données : pour $\beta = 1$, la distribution de Tweedie s'apparente à une distribution de Poisson, pour $\beta = 0$, c'est une distribution gamma (ou exponentielle). Dans le cas où $\beta = 2$, cela correspond à la distribution de la loi normale (ou loi de Gauss) de moyenne μ et de variance σ^2 :

$$p(x|\theta,\lambda) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \tag{3.12}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{\sigma^2} d_{\phi_2}(x|\mu)\right) \tag{3.13}$$

avec $g(x,\lambda) = \frac{1}{\sqrt{2\pi\sigma^2}}$, $\lambda = \sigma^2$ et $d_{\phi_2}(x|\mu)$ qui correspond alors à la distance Euclidienne. Ces distances et divergences, déduites de la distribution de Tweedie, sont alors regroupées dans une sous-classe des divergences de Bregman, appelée β-divergences [Hennequin *et al.*, 2011b]. C'est cette famille de divergences qui est la plus couramment utilisée dans le cadre de la NMF. La Figure 3.2 permet d'illustrer le comportement de ces divergences. Dans le cas où $\beta \in [1,2]$, on constate que d_{ϕ_β} est strictement convexe. En dehors de cet intervalle, les divergences présentent également une partie concave.

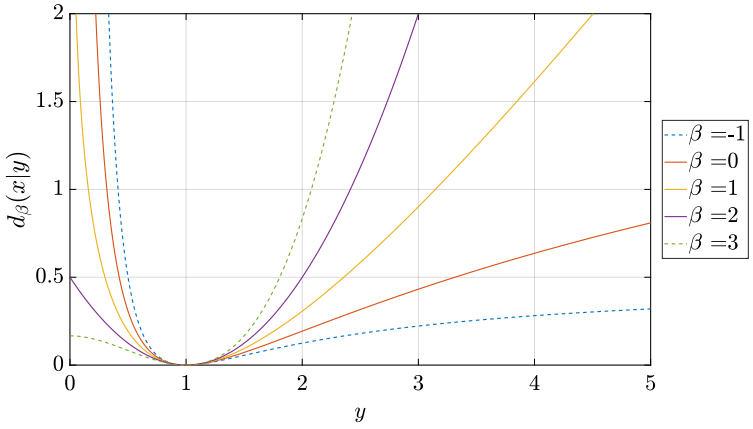


FIGURE 3.2 – Évolution des β-divergences pour un cas simple ($x = 1$).

Dans le reste du document, par souci de clarté, les β-divergences d_{ϕ_β} seront dénommées d_β. Trois β-divergences associées à des valeurs spécifiques de $\beta = [0,1,2]$ et à des distributions particulières, sont le plus souvent utilisées dans le cadre de la NMF et sont présentées dans les parties suivantes.

3.3.1 Distance Euclidienne

Lorsque $\beta = 2$, la β-divergence devient **la distance Euclidienne** (abrégé EUC) :

$$d_2(x|y) = \frac{1}{2}(x-y)^2. \tag{3.14}$$

Cette métrique équivaut à une mesure de similarité entre les points x et y et se révèle très sensible aux grandes variations entre eux en raison de la présence de la puissance carré. En plus des propriétés des divergences de Bregman, la distance Euclidienne en possède 2 autres :

1. **Symétrie** : $d_2(x|y) = d_2(y|x)$.

2. **Inégalité triangulaire** : $d_2(x|y) \leq d_2(x|z) + d_2(x|z)$.

Comme chaque mesure de distance au point (x, y) possède le même poids que les autres couples de points, il est possible, pour la distance Euclidienne, de considérer une pondération $g_{v_{fn}}$ sur les distances selon un critère psycho-acoustique, liée à l'énergie de v_{fn}. Cette pondération permet de prendre en compte dans l'erreur de reconstruction, les points temps-fréquences de faibles énergies [Virtanen, 2004] :

$$d'_2(v_{fn}|w_{fk}h_{kn}) = g_{v_{fn}} d_2(v_{fn}|w_{fk}h_{kn}). \tag{3.15}$$

3.3.2 Divergence de Kullback-Leibler

L'expression 3.11c correspond à **la divergence de Kullback-Leibler** (abrégé K-L) [Kompass, 2007, Cichocki *et al.*, 2006c] . Elle traduit comment une densité de probabilité **x** diverge d'une seconde distribution **y**. En d'autres termes, c'est une mesure de l'information perdue lorsque **y** est utilisée pour approximer **x**. Ne respectant pas les propriétés de symétrie et d'inégalité triangulaire, elle n'est donc pas une distance. La divergence K-L est un cas remarquable puisqu'elle appartient à la fois aux divergences de Bregman et à une autre famille de divergences, les divergences de Csiszar [Cichocki *et al.*, 2006a].

3.3.3 Divergence d'Itakura-Saito

L'expression 3.11b est celle de **la divergence d'Itakura-Saito** (abrégé I-S) [Itakura, 1968, Bertin, 2009]. Cette divergence est la seule des β-divergences à posséder la propriété d'*invariance d'échelle* :

$$d_0(\lambda x|\lambda y) = d_0(x|y). \tag{3.16}$$

Ce rapport signifie que le même poids est attribué entre les fortes et les faibles valeurs de **x**. Ainsi, la divergence aux faibles puissances sonores entre **V** et **WH** aura la même importance que dans les fortes puissances. La relation 3.16 est généralisable à l'ensemble des β-divergences,

$$d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y). \tag{3.17}$$

Dans le cas où $\beta > 0$, la divergence est plus influencée par les composantes de fortes amplitudes. À l'inverse, dans le cas où $\beta < 0$, les composantes ayant une faible amplitude ont une plus grande prépondérance. Cette particularité, d'après [Févotte *et al.*, 2009], est intéressante dans le cadre de signaux audio, et notamment celui des signaux audio musicaux qui possèdent une forte dynamique et dont la puissance décroit exponentiellement en fonction de la fréquence. Une faible valeur de β permet alors de mieux prendre en compte les composantes de moindres amplitudes dans la reconstruction du signal.

3.3.4 Autres familles de divergences

Si on s'est attardé à décrire longuement les β-divergences, d'autres familles de divergences peuvent être utilisées pour la NMF notamment celles appartenant aux divergences de Csiszar (ou appelées f-divergences). Cette famille est toutefois moins utilisée que celle de Bregman. Elle se définit comme

$$D_{\Psi}(\mathbf{x}\|\mathbf{y}) = \mathbf{x}\boldsymbol{\Psi}\left(\frac{\mathbf{x}}{\mathbf{y}}\right). \tag{3.18}$$

Elle intègre notamment la distance de variation totale ($\psi(x) = \frac{1}{2}|x-1|$), la divergence χ^2 ($\psi(x) = (x-1)^2$), la distance d'Hellinger ($\psi(x) = (\sqrt{x}-1)^2$), les divergences α d'Amari ($\psi(x) = \frac{4}{1-\alpha^2}\left(1 - x^{(1+\alpha)/2}\right)$) et la divergence de Kullback-Leibler ($\psi(x) = x\log(x)$). Elle possède plusieurs propriétés (non-négativité, unicité de la solution, symétrie, convexité, dépendance) qui sont détaillées dans [Csiszár *et al.*, 2004]. Enfin, une proposition de généralisation des familles de divergences peut être trouvée dans [Cichocki *et al.*, 2011] où les auteurs généralisent les α et les β-divergences à travers l'équation 3.19 :

$$D_{a,b}(x|y) = -\frac{1}{ab}\left(x^a y^\beta - \frac{a}{a+b}x^{a+b} - \frac{b}{a+b}y^{a+b}\right). \tag{3.19}$$

Les valeurs des coefficients a et b permettent alors d'obtenir soit des β-divergences ($a = 1$) ou des α-divergences ($a+b = 1$) mais également de nouvelles divergences. L'intérêt est d'étendre les familles et les divergences, afin de déterminer de nouveaux algorithmes de mise à jour dont les convergences peuvent être plus rapides, et d'offrir de nombreuses fonctions coûts pouvant mieux s'adapter au problème initial rencontré. Dans le cadre de ce travail, nous nous restreignons aux β-divergences en raison de leur popularité et des nombreux travaux dans la littérature qui s'y réfèrent.

Notons enfin qu'il n'est pas nécessaire de se restreindre aux divergences bien connues : [Vincent *et al.*, 2010] a, par exemple, utilisé la NMF dans le cadre de la transcription de signaux musicaux et a déterminé un résultat optimal pour $\beta = 0{,}5$.

3.4 Mise à jour des formes de W et de H

La minimisation de la β-divergence entre $\mathbf{V}$ et $\mathbf{WH}$ se résout par un processus d'optimisation qui consiste à faire évoluer la forme des matrices $\mathbf{W}$ et $\mathbf{H}$ itérativement à l'aide d'algorithmes de mise à jour qui dépendent du choix de β. Sont présentés ici les algorithmes multiplicatifs les plus couramment utilisés qui garantissent que la contrainte de non-négativité soit respectée.

3.4.1 Algorithme heuristique par descente de gradient

Dans leur premier article consacré à la NMF, Lee et Seung [Lee et Seung, 1999] proposent une première formulation des algorithmes de mises à jour, sans toutefois expliciter leur origine,

pour $\beta \in \{1, 2\}$:

$$\mathbf{W}^{(i+1)} \leftarrow \mathbf{W}^{(i)} \otimes \frac{\left[\left(\mathbf{W}^{(i)}\mathbf{H}\right)^{.(\beta-2)} \otimes \mathbf{V}\right]\mathbf{H}^T}{\left[\mathbf{W}^{(i)}\mathbf{H}\right]^{.(\beta-1)}\mathbf{H}^T}, \tag{3.20a}$$

$$\mathbf{H}^{(i+1)} \leftarrow \mathbf{H}^{(i)} \otimes \frac{\mathbf{W}^T\left[\left(\mathbf{W}\mathbf{H}^{(i)}\right)^{.(\beta-2)} \otimes .\mathbf{V}\right]}{\mathbf{W}^T\left[\mathbf{W}\mathbf{H}^{(i)}\right]^{.(\beta-1)}}. \tag{3.20b}$$

Les termes $A \otimes B$ et $\frac{A}{B}$ sont des produits de Hadamard (respectivement multiplication et division terme à terme). La minimisation de la fonction 3.3 se fait alors alternativement en raison de la propriété de convexité de la β-divergence : pour $\mathbf{W}$ fixé, $\mathbf{H}$ est mis à jour, puis $\mathbf{H}$ est fixé et c'est $\mathbf{W}$ qui est mis à jour. L'obtention des expressions 3.20 et la démonstration que les algorithmes 3.20 permettent de miminiser l'équation 3.3 sont réalisées dans [Lee et Seung, 2000] à l'aide de la méthode de descente de gradient. Cette méthode consiste à faire « glisser » une solution temporaire le long de la pente négative d'une fonction $f(x)$ afin de converger vers la solution [Kivinen et Warmuth, 1994] :

$$x^{(i+1)} \leftarrow x^{(i)} - \eta^{(i)}\nabla f(x^{(i)}) \tag{3.21}$$

avec $\eta^{(i)}$ le pas d'apprentissage. Dans le cadre de la NMF, pour la distance EUC, l'équation 3.21 deviennent :

$$\mathbf{W}^{(i+1)} \leftarrow \mathbf{W}^{(i)} + \eta_{\mathbf{W}}\left[\left(\mathbf{V}\mathbf{H}^{\mathbf{T}}\right)^{(i)} - \left(\mathbf{W}\mathbf{H}\mathbf{H}^{\mathbf{T}}\right)^{(i)}\right], \tag{3.22a}$$

$$\mathbf{H}^{(i+1)} \leftarrow \mathbf{H}^{(i)} + \eta_{\mathbf{H}}\left[\left(\mathbf{W}^{\mathbf{T}}\mathbf{V}\right)^{(i)} - \left(\mathbf{W}^{\mathbf{T}}\mathbf{W}\mathbf{H}\right)^{(i)}\right] \tag{3.22b}$$

avec $\eta_{\mathbf{W}} = \frac{\mathbf{W}}{\mathbf{W}\mathbf{H}\mathbf{H}^{\mathbf{T}}}$ et $\eta_{\mathbf{H}} = \frac{\mathbf{H}}{\mathbf{W}^{\mathbf{T}}\mathbf{W}\mathbf{H}}$, les pas d'apprentissages respectifs choisis judicieusement afin d'obtenir les algorithmes de mises à jour 3.20. Cette méthode est également développée pour la divergence de Kullback-Leibler dans le même article mais n'est alors pas étendu à l'ensemble de la famille des β-divergences. Leur proposition n'apporte alors pas de preuve directe de la convergence des algorithmes vers la solution du problème posé. Cette preuve sera obtenue à l'aide de l'algorithme de *majorisation-minimisation*.

3.4.2 Algorithme multiplicatif par *majorisation-minimisation*

Une seconde approche [Cichocki *et al.*, 2006b] consiste à exprimer le gradient de la fonction de coût, $\nabla_x D(x)$, comme la différence entre deux fonctions non-négatives :

$$\nabla_x D(x) = \nabla_x^+ D(x) - \nabla_x^- D(x). \tag{3.23}$$

La fonction de coût $D(x)$ est alors minimisée lorsque, pour un point donné, le gradient est nul ($\nabla_x^+ D(x) = \nabla_x^- D(x)$). Dans le cas de la NMF, les équations de mise à jour de $\mathbf{W}$ et $\mathbf{H}$

deviennent :

$$\mathbf{W}^{(i+1)} \leftarrow \mathbf{W}^{(i)} \frac{\nabla^{-}_{\mathbf{W}} D(\mathbf{V}\|\mathbf{WH})}{\nabla^{+}_{\mathbf{W}} D(\mathbf{V}||\mathbf{WH})}, \tag{3.24a}$$

$$\mathbf{H}^{(i+1)} \leftarrow \mathbf{H}^{(i)} \frac{\nabla^{-}_{\mathbf{H}} D(\mathbf{V}\|\mathbf{WH})}{\nabla^{+}_{\mathbf{H}} D(\mathbf{V}\|\mathbf{WH})}. \tag{3.24b}$$

En considérant des valeurs initiales positives ou nulles dans $\mathbf{W}$ et $\mathbf{H}$, le processus garantit la non-négativité des valeurs itérées. Par cette approche, la minimisation de la fonction de coût de l'équation 3.3 a, dans un premier temps, été observée sans toutefois être démontrée. Kompass [Kompass, 2007] propose une preuve de son efficacité pour $\beta \in \{1, 2\}$ en considérant une fonction auxiliaire majorante à minimiser. Cette approche est la base de l'algorithme de *majorisation-minimisation* que Févotte et Idier [Févotte et Idier, 2011] ont étendu à l'ensemble des β-divergences.

3.4.2.1 Définition de la fonction auxiliaire

Pour réaliser une fonction auxiliaire, plusieurs conditions sont à considérer :

— la mise à jour d'une des deux matrices se fait pour l'autre matrice fixée.

— Comme l'approximation de la NMF est transposable ($\mathbf{V} \approx \mathbf{WH} \Leftrightarrow \mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$), les mises à jour de $\mathbf{W}$ et de $\mathbf{H}$ sont équivalentes (à cette transposition près).

— Comme le problème de la minimisation peut se restreindre à chaque composante n (équation 3.2), il est possible de limiter l'étude au cas de la mise à jour de $\mathbf{h}$, un vecteur colonne n issu de $\mathbf{H}$, pour $\mathbf{W}$ fixé.

Considérons le problème suivant :

$$\min_{\mathbf{h} > \mathbf{0}} C(\mathbf{h}) = D(\mathbf{v}||\mathbf{Wh}) \tag{3.25}$$

avec $\mathbf{W}$ fixé et $\mathbf{v}$ défini à l'équation 3.2. On définit alors la fonction auxiliaire $G(\mathbf{h}|\mathbf{h})$ de $C(\mathbf{h})$ telle que :

$$C(\mathbf{h}^{(i)}) = G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) \quad \forall\ \mathbf{h} \in \mathbb{R}^{+}_{K}, \tag{3.26a}$$

$$C(\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i+1)}) \quad \forall\ \mathbf{h} \in \mathbb{R}^{+}_{K}. \tag{3.26b}$$

La détermination d'un $\mathbf{h}$ optimal est alors réalisée par un processus itératif afin que

$$\mathbf{h}^{(i+1)} = \arg\min_{\mathbf{h} \geq 0} G(\mathbf{h}|\mathbf{h}^{(i)}), \tag{3.27a}$$

$$C(\mathbf{h}^{(i+1)}) \leq G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) = C(\mathbf{h}^{(i)}). \tag{3.27b}$$

La condition 3.27b permet alors d'obtenir une valeur itérée qui génère un algorithme monotone, c'est-à-dire qu'il certifie la diminution de la fonction de coût à chaque itération ($C(\mathbf{h}(i+1)) < C(\mathbf{h}(i))$. La minimisation de $G(\mathbf{h})$, supposée plus simple, permet, par extension, celle de

TABLEAU 3.1 – Fonctions concaves, convexes et constantes selon β.

	$\beta < 1$ et $\beta \neq 0$	$\beta = 0$	$1 \leq \beta \leq 2$	$\beta > 2$
$\breve{d}(x\|y)$	$-\frac{1}{\beta-1}xy^{\beta-1}$	xy^{-1}	$d_\beta(x\|y)$	$\frac{1}{\beta}y^\beta$
$\breve{d}'(x\|y)$	$-xy^{\beta-2}$	$-xy^{-2}$	$d'_\beta(x\|y)$	$y^{\beta-1}$
$\hat{d}(x\|y)$	$\frac{1}{\beta}y^\beta$	$\log y$	0	$\frac{1}{\beta-1}xy^{\beta-1}$
$\hat{d}'(x\|y)$	$y^{\beta-1}$	y^{-1}	0	$-xy^{\beta-2}$
$\bar{d}(x\|y)$	$\frac{1}{\beta(\beta-1)}x^\beta$	$x(\log x - 1)$	0	$\frac{1}{\beta(\beta-1)}x^\beta$

$C(\mathbf{h})$. La preuve de la convergence d'un algorithme est présente lorsqu'une suite d'itérations successives tend vers un point $\mathbf{h}^*$ qui satisfait les conditions de Karush-Kuhn-Tucker [Févotte et Idier, 2011, Kuhn, 1976] :

$$\min\left\{\mathbf{h}^*, \nabla_{\mathbf{h}} C(\mathbf{h}^*)\right\} = \mathbf{0}_K \tag{3.28}$$

où $\mathbf{0}_K$ est un vecteur nul. En satisfaisant la condition 3.28, l'algorithme est dit convergent puisque la série de valeurs itérées $\mathbf{h}^(i)$ converge vers un point limite $\mathbf{h}^*$.

3.4.2.2 Construction de la fonction auxiliaire

L'équation 3.25 peut s'exprimer sous la forme $C(\mathbf{h}) = \sum_f d_\beta\left(v_f|\left[\mathbf{Wh}\right]_f\right)$ avec la divergence $d_\beta(x|y)$ qui se décompose comme une somme d'une fonction convexe, $\breve{d}(x|y)$, concave, $\hat{d}(x|y)$ et constante, $\bar{d}(x)$ dont les valeurs sont détaillées dans le Tableau 3.1 :

$$d_\beta(x|y) = \breve{d}(x|y) + \hat{d}(x|y) + \bar{d}(x). \tag{3.29}$$

Dans le cas où $\beta \in \{1, 2\}$, la partie concave et constante sont nulles, ce qui se vérifie dans la Figure 3.2. La fonction auxiliaire majorante $G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)})$ est obtenue en majorant ces trois parties séparément : par une inégalité de Jensen pour la partie convexe et par une approximation de Taylor au premier ordre (qui équivaut à sa tangente) pour la partie concave. La fonction auxiliaire $G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)})$ devient alors

$$\begin{aligned} G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) = \sum_f \left[\sum_k \frac{w_{fk}h_k^{(i)}}{\tilde{v}_f}\breve{d}\left(v_f|\tilde{v}_f\frac{h_k^{(i+1)}}{h_k^{(i)}}\right)\right] \\ + \left[\hat{d}'(v_f|\tilde{v}_f)\sum_k w_{fk}(h_k^{(i+1)} - h_k^{(i)}) + \hat{d}(v_f|\tilde{v}_f)\right] + \bar{d}(v_f) \end{aligned} \tag{3.30}$$

avec $\mathbf{h}^{(i+1)}$, le vecteur $\mathbf{h}$ à mettre à jour, $\mathbf{h}^{(i)}$, le vecteur actuel de $\mathbf{h}$, $\tilde{v}_f = \left[\mathbf{Wh}^{(i)}\right]_f$. La fonction 3.30 est alors minimisée en déterminant le zéro de sa dérivée selon h_k :

$$\nabla_{h_k} G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) = \sum_f w_{fk}\left[\breve{d}'\left(v_f|\tilde{v}_f\frac{h_k^{(i+1)}}{h_k^{(i)}}\right) + \hat{d}'(v_f|\tilde{v}_f)\right]. \tag{3.31}$$

De l'équation 3.31, l'expression de $h_k^{(i+1)}$ est déterminée :

$$h_k^{(i+1)} = h_k^{(i)} \left(\frac{\sum_f w_{fk} v_f \tilde{v}_f^{(\beta-2)}}{\sum_f w_{fk} \tilde{v}_f^{(\beta-1)}} \right)^{\gamma(\beta)}, \tag{3.32}$$

$$= h_k^{(i)} \left(\frac{\nabla_{h_k}^- C(\tilde{\mathbf{h}})}{\nabla_{h_k}^+ C(\tilde{\mathbf{h}})} \right)^{\gamma(\beta)} \tag{3.33}$$

avec

$$\gamma(\beta) = \begin{cases} \dfrac{1}{2-\beta}, & \beta < 1, \quad (3.34a) \\ 1, & 1 \leq \beta \leq 2, \quad (3.34b) \\ \dfrac{1}{\beta-1}, & \beta > 2. \quad (3.34c) \end{cases}$$

L'algorithme 3.32 déduit est similaire à l'algorithme heuristique à descente de gradient (équation 3.20) et ne diffère que par la présence de l'exposant $\gamma(\beta)$. Pour $\beta \in \{1, 2\}$, où $d_\beta(x|y)$ est strictement convexe, les deux algorithmes sont même égaux. En dehors de cet intervalle, l'algorithme de *majorisation-minimisation* amène la preuve de la décroissance de l'équation 3.25 pour tout β ce qui n'était qu'observé avec l'algorithme heuristique initial. Ce procédé peut être étendu à $\mathbf{W}$, avec comme fonction auxiliaire $K(\mathbf{w}^{(i+1)}|\mathbf{w}^{(i)})$:

$$K(\mathbf{w}^{(i+1)}|\mathbf{w}^{(i)}) = \sum_f \left[\sum_k \frac{w_{fk}^{(i+1)} h_k}{\tilde{v}_f} \breve{d}\left(v_f | \tilde{v}_f \frac{w_{fk}^{(i+1)}}{w_{fk}^{(i)}} \right) \right] + \left[\grave{d}(v_f|\tilde{v}_f) \sum_k (w_{fk}^{(i+1)} - w_{fk}^{(i)}) h_k + \hat{d}(v_f|\tilde{v}_f) \right] + \bar{d}(v_f). \tag{3.35}$$

L'expression de w_{fk} est alors déduite :

$$w_{fk}^{(i+1)} \leftarrow w_{fk}^{(i)} \left(\frac{h_k v_f \tilde{v}_f^{(\beta-2)}}{h_k \tilde{v}_f^{(\beta-1)}} \right)^{\gamma(\beta)}. \tag{3.36}$$

Les expressions 3.32 et 3.36, généralisées sous formes matricielles, donnent alors les expressions 3.37.

$$\mathbf{W}^{(i+1)} \leftarrow \mathbf{W}^{(i)} \otimes \left(\frac{\left[\left(\mathbf{W}^{(i)} \mathbf{H} \right)^{\cdot(\beta-2)} \otimes \mathbf{V} \right] \mathbf{H}^T}{\left[\mathbf{W}^{(i)} \mathbf{H} \right]^{\cdot(\beta-1)} \mathbf{H}^T} \right)^{\gamma(\beta)}, \tag{3.37a}$$

$$\mathbf{H}^{(i+1)} \leftarrow \mathbf{H}^{(i)} \otimes \left(\frac{\mathbf{W}^T \left[\left(\mathbf{W} \mathbf{H}^{(i)} \right)^{\cdot(\beta-2)} \otimes \mathbf{V} \right]}{\mathbf{W}^T \left[\mathbf{W} \mathbf{H}^{(i)} \right]^{\cdot(\beta-1)}} \right)^{\gamma(\beta)}. \tag{3.37b}$$

C'est ainsi l'algorithme de *majorisation-minimisation* qui est implémenté pour réaliser la NMF dans les parties 5.5 et 6.3.

3.4.3 Autres approches

D'autres algorithmes ont été proposés à partir d'approches différentes comme la méthode des moindres carrés alternés [Cichocki et Zdunek, 2007, Berry *et al.*, 2007] qui consiste à minimiser successivement la distance EUC entre $\mathbf{V}$ et $\mathbf{WH}$ en fixant chaque variable alternativement,

$$\mathbf{W}^{(i+1)} = \arg\min_{\mathbf{W}>0} D\left(\mathbf{V}||\mathbf{W}^{(i)}\mathbf{H}^{(i)}\right), \tag{3.38a}$$

$$\mathbf{H}^{(i+1)} = \arg\min_{\mathbf{H}>0} D\left(\mathbf{V}||\mathbf{W}^{(i+1)}\mathbf{H}^{(i)}\right). \tag{3.38b}$$

Pour résoudre les équations 3.38, Zdunek et Cichocki [Zdunek et Cichocki, 2006] proposent d'utiliser la méthode de Newton alors que Lin [Lin, 2007] utilise la méthode par projection de gradient. Si ces méthodes offrent des convergences plus rapides que les algorithmes multiplicatifs, elles sont alourdies par la présence de matrice hessienne et ne permettent pas d'intégrer aussi facilement des contraintes sur $\mathbf{W}$ ou $\mathbf{H}$. De plus, ces méthodes de résolutions ne sont adaptées que pour la distance EUC ou la divergence K-L, ce qui restreint les possibilités d'adapter les fonctions coûts aux différents problèmes. En conséquence, bien que plus lent que l'algorithme des Moindres Carrés Alternés, l'utilisation de l'algorithme *majorisation-minimisation* permet l'utilisation de l'ensemble des β-divergences et d'ajouter plus facilement des contraintes sur les éléments (voir partie 3.8).

3.5 Analyse Probabiliste en Composantes Latentes

Il est à noter qu'une autre approche de la NMF existe à travers un pendant probabiliste : l'Analyse Probabiliste en Composantes Latentes (abrégé PLCA pour *Probabilistic Latent Composent Analysis* en anglais) [Hofmann, 2001, Cazau et Nuel, 2017]. Elle considère l'ensemble des points d'un spectrogramme $V_{F\times N}$ comme le résultat d'un tirage de $F \times N$ variables indépendantes. Cette distribution suit une loi de distribution discrète paramétrique $P_\Lambda(f,n)$ où Λ résume l'ensemble de ces paramètres. En introduisant la variable aléatoire latente (ou cachée) k, on obtient :

$$P_\Lambda(f,n) = \sum_k P(k)\,P(f,n|k), \tag{3.39}$$

$$= \sum_n P(k)P(n|k)\,P(f|k) \tag{3.40}$$

où $P(n|k)$ est assimilée aux activateurs temporels, $P(f|k)$ aux spectres du dictionnaire (appelés atomes) et $P(k)$ est le poids relatif de chaque composante. Les paramètres de la loi de distribution Λ sont obtenus en maximisant la vraisemblance des observations par un algorithme d'Esperance-Maximisation (*Expectation-Maximization* en anglais). Les expressions de mise à

jour de chaque distribution sont disponibles en vue de maximiser la vraisemblance ([Shashanka *et al.*, 2008]) et permet ainsi de vérifier que la PLCA et la NMF sont des approches similaires d'un problème d'approximation [Gaussier et Goutte, 2005].

Plusieurs variantes de la PLCA existent également comme la PLCA invariante par translation (shift-invariant PLCA) [Smaragdis et Raj, 2007] ou la PLCA invariante par changement d'échelle (scale-invariant PLCA) [Hennequin *et al.*, 2011a] qui permet de transposer des spectrogrammes décomposés en échelle invariante (par une transformation en Q-constant) en une échelle linéaire (obtenue par une TFCT par exemple).

3.6 Apprentissage du dictionnaire

L'utilisation de la NMF nécessite deux étapes : une phase d'apprentissage du dictionnaire et une phase de test (Figure 3.3) :

— Durant la phase d'apprentissage, $\mathbf{W}$ et $\mathbf{H}$ sont des matrices dont le contenu est inconnu. Le corpus d'apprentissage est alors soumis à une méthode d'apprentissage permettant de construire le dictionnaire $\mathbf{W}$ (algorithme de clustering *k-means*, NMF...). Dans le cas de la NMF, la matrice d'activation obtenue $\mathbf{H_0}$, propre à ce corpus, est alors rejetée, seul $\mathbf{W}$ est conservé pour l'étape suivante. Si on s'arrête à la première étape d'apprentissage, la NMF peut alors être vue comme un algorithme de *clustering* [Li et Ding, 2006] et de réduction des données grâce à la contrainte imposée sur les dimensions des matrices ($F \times K + K \times N << F \times N$).

— Pour la phase de test, le dictionnaire $\mathbf{W}$ obtenu est utilisé sur un corpus de test avec $\mathbf{H}$, une nouvelle matrice d'activation inconnue, qui est à déterminer.

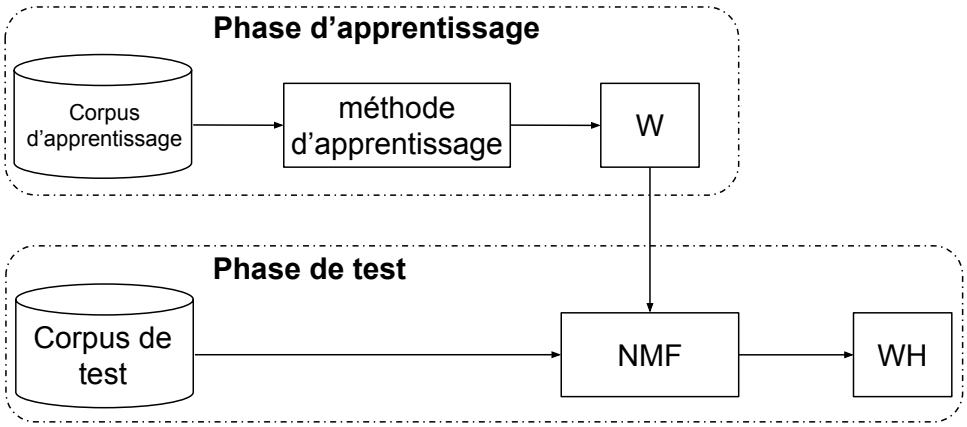


FIGURE 3.3 – Schéma-bloc des étapes de la NMF.

3.6.1 Apprentissage supervisé et non-supervisé

Lorsque les classes de sons du corpus d'apprentissage sont inconnues ou que les échantillons audio sont composés d'un mélange de plusieurs sources, il n'est pas possible de connaitre la classe de son de chaque élément qui constitue $\mathbf{W}$. Toutefois, pour réaliser une séparation de sources, il est nécessaire de classifier les différents éléments entre eux. Une étape de *clustering*, à l'aide d'un algorithme des k-NN (k plus proche voisins) par exemple, est alors nécessaire pour classer ces éléments selon un nombre de catégories défini par l'utilisateur. Ce cas correspond à une *NMF non-supervisée*. À l'inverse, lorsque le jeu de données peut être étiqueté, chaque élément défini dans $\mathbf{W}$ est connu. Ce cas, le plus favorable et le plus simple, correspond à une *NMF supervisée*. En connaissant la classe des éléments de $\mathbf{W}$, la séparation de sources est réalisable. Pour cela, les éléments relatifs à la source d'intérêt i sont extraits soit directement de $\mathbf{W}$ et de $\mathbf{H}$,

$$\tilde{\mathbf{V}}_i = [\mathbf{WH}]_i \,, \tag{3.41}$$

soit cette séparation est réalisée par un filtre de Wiener (ou masquage doux) :

$$\tilde{\mathbf{V}}_i = \frac{[\mathbf{WH}]_i}{\mathbf{WH}} \otimes \mathbf{V}. \tag{3.42}$$

3.6.2 Apprentissage semi-supervisé

L'apprentissage du dictionnaire pose la question de la généralisation des connaissances : comment à partir de données limitées obtenir une NMF efficace sur un ensemble de cas divers et variés ? Cette question se base sur le constat qu'il n'est pas possible de modéliser dans $\mathbf{W}$ l'ensemble des classes de sons qui composent un environnement notamment l'ESU qui est un milieu qui inclut une multitude de sources sonores variables. Pour résoudre ce problème, une des premières solutions est de constituer une base d'apprentissage plus importante que la base de test, en vue d'augmenter la généralisation des connaissances apprises ou leur quantité. Toutefois, cette option n'est parfois pas réalisable soit parce que les données ne sont tout simplement pas disponibles, soit parce que la quantité de données à gérer serait ensuite trop importante et nécessiterait des moyens de calculs puissants pour pouvoir mener à bien l'approximation du signal testé. Une autre possibilité pour tenter de résoudre cette question est de réaliser un apprentissage semi-supervisé, tel que proposé par [Lee *et al.*, 2010, Smaragdis *et al.*, 2007] (Figure 3.4). La NMF semi-supervisée propose de construire un dictionnaire $\mathbf{W}_{F\times(K+J)}$ composé d'éléments appris sur le corpus d'apprentissage, $\mathbf{W_s}$ de dimensions $F \times K$, et d'éléments inconnus, $\mathbf{W_r}$ de dimensions $F \times J$ avec $J << K$. Cette condition est nécessaire afin de focaliser la reconstruction du signal avec les sources présentes dans $\mathbf{W_s}$. L'idée est alors de mettre à jour $\mathbf{W_r}$ lors de la phase de test afin d'y intégrer les autres sources sonores qui ne sont pas apprises dans $\mathbf{W_s}$. On obtient donc

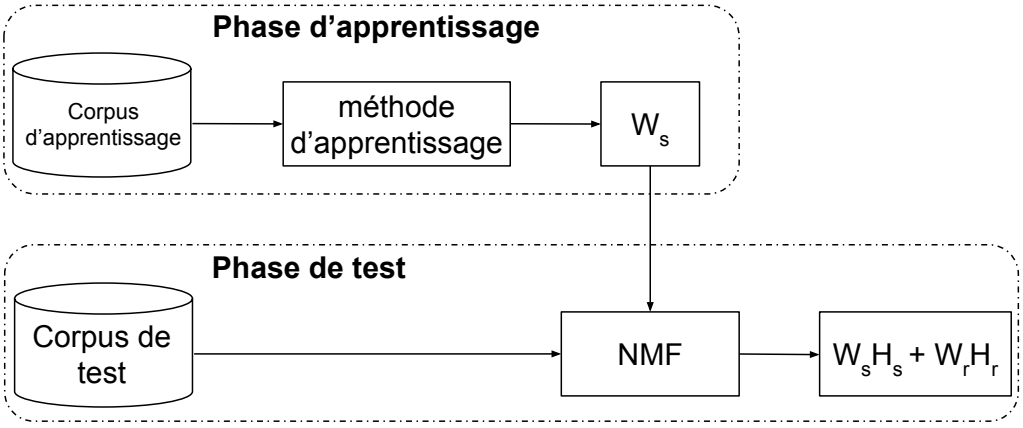


FIGURE 3.4 – Schéma-bloc des étapes de la NMF semi-supervisée.

$$\mathbf{V} \approx \mathbf{WH} = \mathbf{W_s H_s} + \mathbf{W_r H_r} \tag{3.43}$$

avec $\mathbf{W} = [\mathbf{W_s W_r}]$ et respectivement $\mathbf{H} = \begin{bmatrix} \mathbf{H_s} \\ \mathbf{H_r} \end{bmatrix}$ constituée de la matrice $\mathbf{H_s}_{K \times N}$ et de la matrice $\mathbf{H_r}_{J \times N}$. $\mathbf{H_s}$, $\mathbf{W_r}$ et $\mathbf{H_r}$ sont donc les 3 matrices à déterminer lors de la phase de test à l'aide des algorithmes de mises à jour 3.44 [Kitamura *et al.*, 2014] :

$$\mathbf{W_r}^{(i+1)} \leftarrow \mathbf{W_r}^{(i)} \otimes \left(\frac{\left[\left(\mathbf{WH}^{(i)} \right)^{.(\beta-2)} \otimes \mathbf{V} \right] \mathbf{H_r}^T}{\left(\mathbf{WH}^{(i)} \right)^{.(\beta-1)} \mathbf{H_r}^T} \right)^{\gamma(\beta)} \tag{3.44a}$$

$$\mathbf{H_r}^{(i+1)} \leftarrow \mathbf{H_r}^{(i)} \otimes \left(\frac{\mathbf{W_r}^T \left[\left(\mathbf{WH}^{(i)} \right)^{.(\beta-2)} \otimes \mathbf{V} \right]}{\mathbf{W_r}^T \left(\mathbf{WH}^{(i)} \right)^{.(\beta-1)}} \right)^{\gamma(\beta)} \tag{3.44b}$$

$$\mathbf{H_s}^{(i+1)} \leftarrow \mathbf{H_s}^{(i)} \otimes \left(\frac{\mathbf{W_s}^T \left[\left(\mathbf{WH}^{(i)} \right)^{.(\beta-2)} \otimes \mathbf{V} \right]}{\mathbf{W_s}^T \left(\mathbf{WH}^{(i)} \right)^{.(\beta-1)}} \right)^{\gamma(\beta)} . \tag{3.44c}$$

Dans le cas où $\mathbf{W_s}$ est composé de spectres relatifs au trafic, le produit $\mathbf{W_r H_r}$ est supposé inclure des éléments qui appartiennent à d'autres classes de sons. L'extraction du signal trafic est réalisée à partir de $\mathbf{W_s}$ et de $\mathbf{H_s}$,

$$\tilde{\mathbf{V}}_{trafic} = \mathbf{W_s H_s} \tag{3.45}$$

Plusieurs applications de cette méthode ont été proposées pour des tâches de séparation de sources dans des signaux musicaux [Smaragdis *et al.*, 2007] ou pour débruiter des signaux contenant de la voix [Mysore et Smaragdis, 2011, Duan *et al.*, 2012]). Dans [Lefevre *et al.*, 2012], une NMF semi-supervisée est réalisée en contraignant la prépondérance des éléments appris dans

la reconstruction du signal afin de faciliter l'annotation pour réaliser de la séparation de sources dans des signaux musicaux.

3.7 NMF initialisée seuillée

Afin de répondre au problème de généralisation de $\mathbf{W}$, une autre approche est proposée : la *NMF initialisée seuillée* (abrégé NMF IS, à ne pas confondre avec la divergence IS qui est la divergence d'Itakura Saito). Un dictionnaire initial, $\mathbf{W_0}$, est appris sur la source sonore cible (le trafic routier), puis est mis à jour avec $\mathbf{H}$ lors de la phase de test. Cette technique permet d'orienter les mises à jour de $\mathbf{W_0}$ vers la source d'intérêt et ainsi de considérer les connaissances obtenues *a priori* sur la source tout en adaptant le dictionnaire à la scène sonore testée. Après N itérations, un dictionnaire $\mathbf{W'}$ est obtenu, unique à chaque scène. Pour estimer le signal *trafic*, chaque élément k du dictionnaire, $\mathbf{W'}$, est comparé à son spectre d'origine dans $\mathbf{W_0}$ afin de déterminer s'il est encore assimilable à un spectre *trafic*. La comparaison des 2 éléments est réalisée à travers le calcul de leur similarité cosinus :

$$D_\theta(\mathbf{W_0}\|\mathbf{W'}) = \frac{\mathbf{W_0}.\mathbf{W'}}{\|\mathbf{W_0}\|.\|\mathbf{W'}\|} \tag{3.46}$$

qui détermine le cosinus de l'angle θ formé entre les vecteurs $\mathbf{w_0}$ et $\mathbf{w'}$ par le rapport entre leur produit scalaire et leur norme. Dans la suite du document, on se référera à cette distance sous l'abréviation D_θ. Cette métrique est invariante d'échelle et est normée entre -1 et 1 :

- si $D_\theta = 1$, les deux vecteurs sont strictement identiques, $\mathbf{w'}$ est alors considéré comme un élément *trafic*,
- si $D_\theta = 0$, les deux vecteurs sont orthogonaux, $\mathbf{w'}$ n'est pas une élément *trafic* et est donc rejeté,
- si $D_\theta = -1$, les deux vecteurs sont opposés. Ce cas n'est toutefois pas possible en raison de la contrainte de non-négativité.

La similarité entre $\mathbf{W_0}$ et $\mathbf{W'}$ correspond alors à une suite de valeurs comprises entre 0 et 1 puis représentées à travers 2 fonctions (Figure 3.5) :

- une fonction linéaire, $f_{LIN}(k) = D_\theta$,
- une fonction sigmoïde, $f_{SIG}(k) = {}^1/_{(1+\exp(-\lambda D_\theta)}$ avec λ le paramètre d'inflexion de la fonction. Cet opérateur réduit la fenêtre de variation de la distance D_θ, en diminuant les valeurs élevées et en augmentant les valeurs proches de 0.

L'extraction du signal *trafic* est réalisée en pondérant alors le produit $\mathbf{W'H}$ tel que

$$\tilde{\mathbf{V}}_{trafic} = \alpha \otimes [\mathbf{W'H}] \tag{3.47}$$

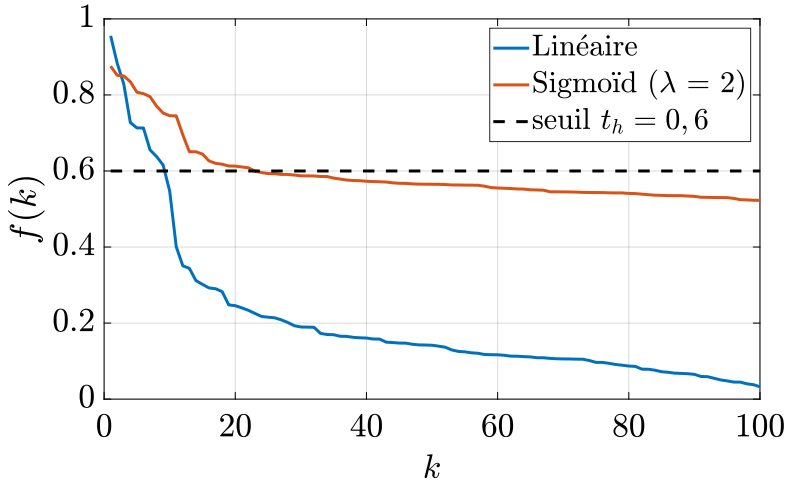


FIGURE 3.5 – Similarité cosinus pour une représentation linéaire et sigmoïdienne de la distance $D_\theta(\mathbf{W_0}\|\mathbf{W}')$ trié dans l'ordre décroissant avec un seuillage dur $t_h = 0{,}6$.

avec $\alpha = [\alpha_1, \alpha_2 \dots \alpha_K]$ qui est défini à partir du calcul de similarité (équation 3.46) et d'une méthode de seuillage. 2 méthodes de seuillages sont envisagées :

— le seuillage dur (*hard thresholding*) [Donoho et Johnstone, 1994] qui consiste à ne considérer dans $\mathbf{W}_{trafic}$ que les éléments de $\mathbf{W}'$ dont la similarité cosinus est supérieure à une valeur seuil t_h :

$$\alpha_k = \begin{cases} 0 & \text{si} \quad f(k) \leq t_h, & \text{(3.48a)} \\ 1 & \text{si} \quad f(k) > t_h, & \text{(3.48b)} \end{cases}$$

— le seuillage *firm* [Fornasier et Rauhut, 2008] qui consiste à pondérer les éléments situés entre deux seuils $t_{f,1}$ et $t_{f,2}$ avec $t_{f,1} < t_{f,2}$, en les normalisant entre 1 et 0 :

$$\alpha_k = \begin{cases} 0, & \text{si } f(k) \leq t_{f,1}, & \text{(3.49a)} \\ \|f(k)\|, & \text{si } t_{f,1} < f(k) \leq t_{f,2}, & \text{(3.49b)} \\ 1, & \text{si } f(k) > t_{f,2} & \text{(3.49c)} \end{cases}$$

avec $\|f(k)\| = \frac{f(k) - \min(f(k))}{\max(f(k)) - \min(f(k))}$

L'allure des pondérations α pour le seuillage dur et *firm* est résumée en Figure 3.6.

On résume les différentes étapes de la NMF IS au travers de l'algorithme 1.

L'intérêt de cette approche est qu'elle permet, par la mise à jour du dictionnaire $\mathbf{W_0}$ de modéliser directement la source sonore *trafic* avec les effets de l'environnement sur la propagation sonore. En effet, la NMF supervisée et semi-supervisée possède un dictionnaire fixe (ou dans une grande partie fixe pour la semi-supervisée) avec lequel elles doivent modéliser l'ensemble de cette source. Par rapport au problème posé dans la partie 1.1, elles contiennent dans $\mathbf{W}$, la source

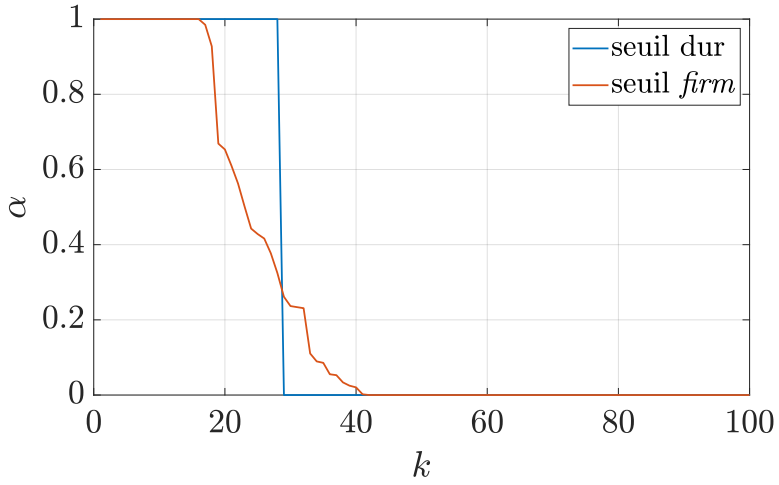


FIGURE 3.6 – Pondération α appliquée à $\mathbf{W}'$, exprimé linéairement, composé de 100 bases avec pour seuil $t_h = 0,2$ (seuillage dur) et $t_{f,1} = 0,15$ et $t_{f,2} = 0,30$ (pour le seuillage *firm*).

Algorithme 1 NMF initialisée seuillée

Initialisation de $\mathbf{W_0}$ sur le corpus d'apprentissage
for i = 1 : nombre itération **do**
  mise à jour de $\mathbf{W_0}$
  mise à jour de $\mathbf{H}$
end for
Calcul de la similarité cosinus $D_\theta(\mathbf{W_0}\|\mathbf{W}')$
if représentation sigmoïde **then**
  $D_\theta(\mathbf{W_0}\|\mathbf{W}') = {}^1/_{(1+\exp(-\lambda D_\theta(\mathbf{W_0}\|\mathbf{W}'))}$
end if
if seuillage *dur* **then**
  α définit selon Eq. 3.48
else if seuillage *firm* **then**
  α définit selon Eq. 3.49
end if
$\tilde{\mathbf{V}}_{trafic} = \alpha\mathbf{W}'\mathbf{H}$

$\hat{s}_{v_j}(f)$. La NMF IS, en calculant $\mathbf{W'}$, contient dans son dictionnaire l'ensemble des sources présentes filtrées par l'environnement urbain. L'extraction du signal *trafic* de $\mathbf{W'H}$ permet alors de déterminer $\hat{S}_{tr.}(f)$, ce qui permet une plus grande généralisation de la méthode. Toutefois, en mettant à jour la forme de chaque élément, on perd la supervision du dictionnaire puisque certains spectres de $\mathbf{W_0}$ appartenant initialement à la source *trafic* peuvent être déviés en une autre source. La méthode de seuillage pour déterminer les composantes *trafic* permet de conserver les éléments qui dévient le moins. Mais cette technique génère un risque de considérer des éléments de la classe *interférante* si le seuil t_h (ou $t_{f,1/2}$) est trop faible ou bien pas assez d'éléments si il est trop élevé.

3.8 NMF avec contraintes

Les différentes variantes de la NMF présentées ne sont soumises, jusqu'ici, qu'à la contrainte de non-négativité avec pour objectif la minimisation de l'équation 3.3. Toutefois, l'ajout de contraintes sur l'apprentissage du dictionnaire ou sur l'allure de la matrice d'activation est possible suivant les connaissances que l'on a *a priori* de ces éléments. Ces contraintes sont alors prises en compte dans la fonction de coût 3.3 par l'ajout d'un second terme pondéré $C(\mathbf{W}, \mathbf{H})$. Le problème devient alors :

$$\min D\left(\mathbf{V} \| \mathbf{WH}\right) + \alpha C(\mathbf{W}, \mathbf{H}) \quad \text{avec} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0. \tag{3.50}$$

Si l'allure de $\mathbf{W}$ ou $\mathbf{H}$ n'est pas celle souhaitée, la fonction de coût 3.50 augmente. Les algorithmes de mise à jour vont alors considérer cette contrainte afin de favoriser les matrices qui permettent de minimiser au mieux cette fonction de coût. Le terme de pondération α permet de faire varier le poids de la contrainte : plus ce terme est grand et plus son influence sera prépondérante. Plusieurs types de contraintes sont décrites et serviront dans la suite de l'étude.

3.8.1 Contrainte de parcimonie

Une des premières contraintes employées est celle de la parcimonie [Hoyer, 2004, Le Roux *et al.*, 2015] : c'est-à-dire l'utilisation d'un nombre réduit d'éléments du dictionnaire à chaque instant t. Cette contrainte renforce la représentation par partie de la NMF en pénalisant les termes qui seraient non nuls. Elle trouve notamment son intérêt dans le cas d'une représentation *sur-complète* du dictionnaire, c'est à dire $K > \max(F, N)$ où l'ajout de la contrainte parcimonieuse permet de réduire la complexité du problème [Eggert et Körner, 2004]. Dans un premier temps, Hoyer [Hoyer, 2004] propose une contrainte telle que

$$C(h_k) = \frac{\sqrt{N} - \left(\sum_{n=1}^{N} |h_{kn}| \sqrt{\sum_{n=1}^{N} h_{kn}^2}\right)}{\sqrt{N} - 1} \tag{3.51}$$

qui équivaut au rapport de la norme ℓ_1 et ℓ_2 normalisée de h_{kn}. Ce rapport est ensuite défini tel que $C(h_k) = C_h$, une valeur constante définie qui fixe la quantité de parcimonie souhaitée

dans $\mathbf{H}$. Un algorithme de gradient de descente permet ensuite, sous cette contrainte, de résoudre l'équation 3.3. Virtanen [Virtanen, 2007] propose l'ajout d'une contrainte $C_{sp}(\mathbf{h})$, le plus souvent considérée comme la norme ℓ_1 des éléments de $\mathbf{H}$,

$$C_{sp}(\mathbf{h}) = \sum_{k=1}^{K} \sum_{n=1}^{N} h_{kn}. \tag{3.52}$$

Cette pénalisation peut facilement être prise en compte dans les algorithmes de *majorisation-minimisation* :

$$\min_{\mathbf{h} > \mathbf{0}} C(\mathbf{h}) = D(\mathbf{v}||\mathbf{W}\mathbf{h}) + \alpha_{sp} C_{sp}(\mathbf{h}) \tag{3.53}$$

avec α_{sp}, la pondération respective à la contrainte de parcimonie. La fonction auxiliaire pénalisée $G_p(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)})$ devient alors

$$G_p(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) = G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) + \alpha_{sp} C_{sp}(\mathbf{h}^{(i+1)}) \tag{3.54}$$

et a pour dérivée selon h_k

$$\nabla_{h_k} G_p(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) = \sum_f w_{fk} \left[\breve{d}' \left(v_f | v_f^{(i)} \frac{h_k^{(i+1)}}{h_k^{(i)}} \right) + \tilde{d}'(v_f | v_f^{(i)}) \right] + \alpha_{sp}. \tag{3.55}$$

L'algorithme de mise à jour de $\mathbf{H}$ devient alors

$$\mathbf{H}^{(i+1)} \leftarrow \begin{cases} \mathbf{H}^{(i)} \otimes \left(\dfrac{\mathbf{W}^T \left[\left(\mathbf{W}\mathbf{H}^{(i)} \right)^{(\beta-2)} \otimes \mathbf{V} \right]}{\mathbf{W}^T \left[\mathbf{W}\mathbf{H}^{(i)} \right]^{(\beta-1)} + \alpha_{sp}} \right)^{\gamma(\beta)} & \beta < 2, \quad (3.56a) \\ \mathbf{H}^{(i)} \otimes \left(\dfrac{\mathbf{W}^T \left[\left(\mathbf{W}\mathbf{H}^{(i)} \right)^{(\beta-2)} \otimes \mathbf{V} \right] - \alpha_{sp}}{\mathbf{W}^T \left[\mathbf{W}\mathbf{H}^{(i)} \right]^{(\beta-1)}} \right)^{\gamma(\beta)} & \beta \geq 2. \quad (3.56b) \end{cases}$$

Dans le cas où $\beta \geq 2$, la contrainte apparait au numérateur et est soustractive. Ce comportement peut être problématique dans le cadre de la non-négativité suivant la value de α_{sp} [Févotte et Idier, 2011]. En Figure 3.7, un exemple de l'effet de la parcimonie sur l'allure d'un élément j de la matrice $\mathbf{H}$ pour 3 valeurs de parcimonie. Avec l'augmentation de la valeur de la pondération, certaines activations présentes lorsque la contrainte est absente, mais de faibles amplitudes, deviennent nulles.

3.8.2 Contrainte de régularité temporelle

La mise à jour de la matrice d'activation $\mathbf{H}$ se fait par défaut trame par trame sans considérer de liens entre les trames n et les précédentes. Néanmoins, la plupart des sons réels ont une

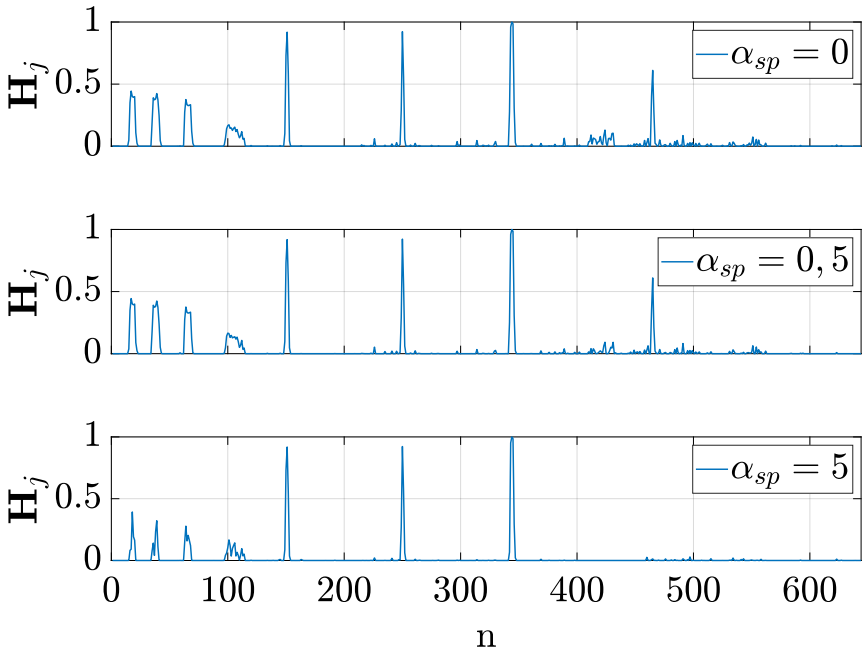


FIGURE 3.7 – Exemple de l'effet de la parcimonie pour une scène du corpus *alerte* pour $\alpha_{sp} \in \{0,\ 0,5,\ 5\}$.

évolution temporelle lente. Prendre en compte l'évolution des trames temporelles adjacentes peut permettre que la forme des activateurs soit plus réaliste et que l'outil soit plus robuste dans la reconstruction du signal. Cette contrainte trouve son intérêt dans la musique ([Virtanen, 2003, Févotte, 2011]) où les instruments (à l'exception des instruments percussifs) peuvent jouer des notes durant, au moins, plusieurs centaines de milli-secondes. Mais c'est aussi le cas au sein d'environnements sonores urbains où les sons (notamment le trafic routier) ont des variations lentes, de plusieurs secondes. L'une des approches les plus citées est celle de Virtanen [Virtanen, 2007], qui fut ensuite généralisée sous la forme d'un algorithme générique dans [Févotte *et al.*, 2018], avec l'ajout d'une contrainte $C_t(\mathbf{H})$:

$$C_t(\mathbf{H}) = \sum_{n=1}^{K} \sum_{n=2}^{N} \left(h_{kn} - h_{k(n-1)} \right)^2 . \tag{3.57}$$

La pondération de cette contrainte peut être constante sur tous les éléments (α_t) ou bien être variable selon k ($\alpha_{t,k}$) et doit donc être placée dans la somme. Par l'ajout de cette contrainte, les fortes variations d'un vecteur d'activation entre l'indice n et $n-1$ sont pénalisées par la mise au *carré* de leur distance. La mise à jour de $\mathbf{H}$ privilégie alors les variations plus lentes pour réduire le poids de la contrainte $C_t(\mathbf{H})$. L'algorithme de mise à jour devient :

$$\mathbf{H}^{(i+1)} \leftarrow \mathbf{H}^{(i)} \otimes \left(\frac{\mathbf{W}^T \left[\left(\mathbf{W}\mathbf{H}^{(i)} \right)^{(\beta-2)} \otimes \mathbf{V} \right] + 2A \otimes \left(\overrightarrow{\mathbf{H}}^{(i)} + \overleftarrow{\mathbf{H}}^{(i)} \right)}{\mathbf{W}^T \left[\mathbf{W}\mathbf{H}^{(i)} \right]^{(\beta-1)} + 2A \otimes \left(\mathbf{H}^{(i)} + \overleftrightarrow{\mathbf{H}}^{(i)} \right)} \right)^{\gamma(\beta)} \tag{3.58}$$

avec

$$A = \begin{bmatrix} \alpha_{t,1} & \cdots & \alpha_{t,1} \\ \alpha_{t,2} & \cdots & \alpha_{t,2} \\ \vdots & \ddots & \vdots \\ \alpha_{t,K} & \cdots & \alpha_{t,K} \end{bmatrix}, \tag{3.59a}$$

$$\overrightarrow{\mathbf{H}} = \begin{bmatrix} 0 & h_{1,1} & h_{1,2} & \cdots & h_{1,N-1} \\ 0 & h_{2,1} & h_{2,2} & \cdots & h_{2,N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & h_{K,1} & h_{K,2} & \cdots & h_{K,N-1} \end{bmatrix}, \tag{3.59b}$$

$$\overleftarrow{\mathbf{H}} = \begin{bmatrix} h_{1,2} & h_{1,3} & \cdots & h_{1,N} & 0 \\ h_{2,2} & h_{2,3} & \cdots & h_{2,N} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{K,2} & h_{K,3} & \cdots & h_{K,N} & 0 \end{bmatrix}, \tag{3.59c}$$

$$\overleftrightarrow{\mathbf{H}} = \begin{bmatrix} 0 & h_{1,2} & \cdots & h_{1,N-1} & 0 \\ 0 & h_{2,2} & \cdots & h_{2,N-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & h_{K,2} & \cdots & h_{K,N-1} & 0 \end{bmatrix}. \tag{3.59d}$$

A résume les valeurs des contraintes selon k, $\overrightarrow{\mathbf{H}}$, les valeurs de $\mathbf{H}$ à l'instant $n-1$, $\overleftarrow{\mathbf{H}}$, les valeurs de $\mathbf{H}$ à l'instant $n+1$ et enfin $\overleftrightarrow{\mathbf{H}}$, les valeurs de $\mathbf{H}$ sur l'intervalle $[2, N-1]$. L'impact du coefficient de pondération α_t sur les activations de $\mathbf{H}$ est représenté en Figure 3.8. Plus cette pondération est forte plus l'allure des activateurs est régulière. C'est cet algorithme qui a été implémenté et testé pour les travaux de cette thèse.

D'autre approches ont également été étudiées. Dans, [Févotte, 2011], dans le cas de la séparation de sources audio et la transcription d'un morceau de musique, une contrainte de régularité est proposée pour la divergence I-S. La particularité de la contrainte apposée est qu'elle se base elle-même sur la divergence I-S : $C_{I-S}(\mathbf{H}) = \sum_{k=1}^{K} \sum_{n=2}^{N} d_0(h_{k(n-1)}|h_{kn})$. L'algorithme de mise à jour est déduit à partir de l'algorithme de *majorisation-minimisation* (voir partie 3.4.2). Dans [Essid et Févotte, 2013], une contrainte, similaire à l'équation 3.57 ($C_{MM}(\mathbf{H}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{n=2}^{N} \left(h_{kn} - h_{k(n-1)} \right)^2$), est proposée dans le cas de la structuration de documents audiovisuels pour une divergence K-L avec, là encore, l'utilisation d'un algorithme de *majorisation-minimisation*. Cette approche a été développée durant la thèse pour la divergence I-S et la distance EUC mais n'a pas été utilisée pour ces travaux. Les détails des calculs se situent en annexe A. Enfin on peut citer [Pascual-Montano *et al.*, 2006] où une contrainte de « non-smoothness »

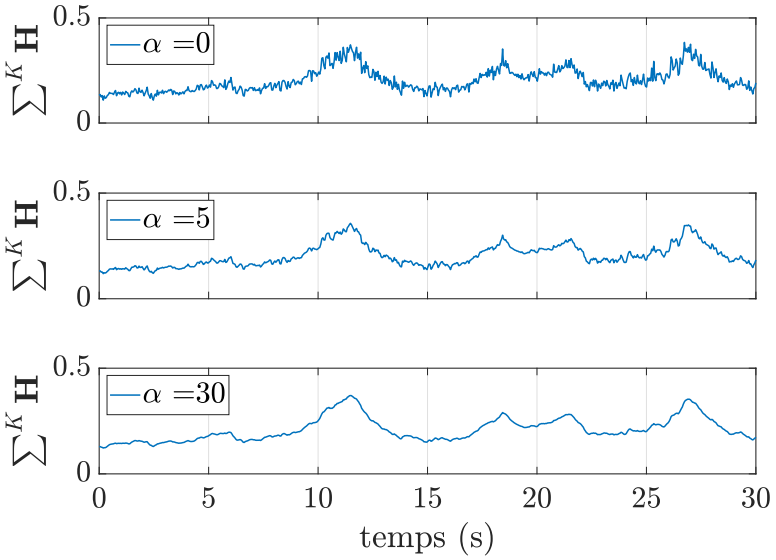


FIGURE 3.8 – Influence de la pondération α de la contrainte temporelle $C_t(\mathbf{H})$ sur la somme ; selon K, des activateurs sur une scène audio de 30 secondes.

est considérée. L'idée est alors d'imposer de l'adoucissement dans la NMF sur une des deux matrices pour générer, en réaction inverse, de la parcimonie dans l'autre matrice. Le problème est ainsi posé :

$$\mathbf{V} \approx \mathbf{WSH} \tag{3.60}$$

où $\mathbf{S}$ est une matrice de *smoothness* telle que

$$\mathbf{S} = (1 - \alpha_t)\mathbf{I} + \frac{\alpha}{K}\mathbf{1}\mathbf{1}^T \tag{3.61}$$

avec $\mathbf{I}$ la matrice identité, $\mathbf{1}$ un vecteur unitaire et α_t un paramètre de *smoothness* compris entre 0 et 1. Lorsque $\alpha = 0$, la régularité temporelle est nulle alors que pour $\alpha_t = 1$, le produit $\mathbf{SH}$ (ou $\mathbf{WS}$) génère un vecteur constant qui correspond à la moyenne des éléments de $\mathbf{H}$ (ou $\mathbf{W}$). Ce cas présente alors une parcimonie nulle puisque toute les entrées sont égales et non-nulles (là où une plus forte parcimonie présente des valeurs proches de 0).

3.8.3 Autres contraintes

D'autres contraintes existent dans la littérature mais ne seront pas étudiées dans ces travaux. On peut citer une contrainte sur l'**harmonicité** des éléments de $\mathbf{W}$ qui a été proposée dans [Vincent *et al.*, 2008] dans le cadre de la transcription d'une mélodie de piano. Le dictionnaire est alors décomposé en une somme de partiels sinusoïdes harmoniques ou inharmoniques pondérées selon une enveloppe spectrale. [Rigaud *et al.*, 2012] propose une autre approche où une contrainte d'inharmonicité est imposée permettant à chaque élément de $\mathbf{W}$ de modifier la fréquence de

chaque partiel tout en contraignant l'ensemble à suivre une loi d'inharmonicité.

La **NMF locale** a été proposée par [Li *et al.*, 2001] qui contraint la fonction de coût avec trois pénalités : i) le nombre de bases K doit être minimisé, ii) les bases doivent être le plus orthogonales que possible, iii) seules les bases qui donnent le plus d'informations sont conservées. Son application à la classification d'instruments de musique dans [Benetos *et al.*, 2006] est toutefois peu efficace.

Dans le cadre de la NMF semi-supervisée, une contrainte est proposée dans [Yagi *et al.*, 2012], puis étendue dans [Kitamura *et al.*, 2014] pour toute valeur de β, dans le cadre de la séparation de signaux musicaux afin de **maximiser la distance** (ou la divergence) entre le dictionnaire appris $\mathbf{W_s}$ et le dictionnaire libre $\mathbf{W_r}$ afin de s'assurer qu'un minimum d'information d'intérêt y soit intégré. Dans [Wang *et al.*, 2016], c'est une contrainte de propagation qui est considérée où un nouveau dictionnaire $\mathbf{W}$ est construit dans lequel les éléments similaires dans $\mathbf{W_r}$ à ceux appris et étiquetés sont pondérés. Dans [Lefevre *et al.*, 2012], la contrainte ajoutée consiste à pondérer le poids des éléments appris dans la reconstruction de $\mathbf{V}$ afin de donner une prépondérance différente des éléments non-appris.

3.9 Conlusion du chapitre

La Factorisation en Matrices Non-négative est une méthode qui permet l'approximation d'un spectrogramme en amplitude (ou en puissance) d'un enregistrement audio par le produit de deux matrices : $\mathbf{W}$, le dictionnaire composé de spectres sonores, et $\mathbf{H}$, la matrice d'activation. Différentes formes de NMF ont été proposées qui diffèrent selon les techniques d'apprentissage de leur dictionnaire (non-supervisée, supervisée et semi-supervisée) ou des contraintes qui y sont appliquées (parcimonie, régularité temporelle). Plusieurs de ces NMF sont implémentées dans le cas de ces travaux et testées en vue de déduire, sur des corpus de sons présentés dans le chapitre suivant, le niveau sonore du trafic routier. La **NMF supervisée**, **semi-supervisée** et, celle proposée, **initialisée seuillée** sont les trois approches retenues. L'influence de la contrainte de **régularité temporelle** $C_t(\mathbf{H})$ est également observée pour un des corpus de sons dans le chapitre 6. Les 3 β-divergences détaillées (**distance EUC** ($\beta = 2$), **divergence K-L** ($\beta = 1$) et **I-S** ($\beta = 0$)) seront celles utilisées sur chaque corpus.

Chapitre 4

Création de corpus de mixtures sonores urbaines

Résumé

Les étapes menant à la création des corpus de sons sur lesquelles la NMF est appliquée sont présentées. Après avoir exposé différentes méthodes pour générer un ESU, l'outil retenu pour ces travaux, le logiciel *SimScene*, est décrit. La formation de la base de données de sons et les enregistrements audio de passages de véhicules sont ensuite détaillés. Le premier corpus créé, appelé corpus d'évaluation *Ambiance* est présenté, celui-ci a pour objectif l'étude du fonctionnement de la NMF. Le second corpus *SOUR* (Scènes sOnores Urbaines Réalistes), dont la construction est basée sur des annotations d'enregistrements audio, est ensuite exposé. Le rendu de ce corpus est évalué par un test perceptif visant à déterminer le réalisme de son rendu.

La méthode de la NMF a été retenue comme estimateur afin d'obtenir les niveaux sonores du trafic routier. Toutefois, nous choisissons de ne pas l'appliquer directement sur des enregistrements audio réalisés en ville. En effet, le trafic sonore y est alors mélangé à d'autres sources sonores. Son niveau sonore exact est alors inconnu. Il n'est donc pas possible de comparer l'estimation fournie par la méthode NMF avec une valeur de référence exacte et ainsi d'évaluer ses performances et sa justesse. L'approche choisie est ici d'appliquer cet estimateur sur des scènes sonores urbaines simulées où la contribution du trafic routier sera connue avec précision et où les estimations des niveaux sonores pourront être comparées aux niveaux exacts. Ce choix soulève plusieurs questions : comment composer des mixtures sonores urbaines aussi réalistes que des enregistrements sonores ? Comment s'assurer que les corpus de sons sur lesquels sont appliqués les algorithmes permettent de tester les limites des estimateurs ? Il en effet nécessaire de s'attarder sur la création des mixtures sonores urbaines et sur la qualité de leur composition afin que

les performances de l'estimateur soit similaires entre les scènes simulées et des enregistrements audio faits en ville.

La création d'environnements sonores urbains réalistes dépasse la question de l'estimation du niveau sonore du trafic. L'intérêt d'utiliser des scènes sonores issus d'un processus de simulation est de permettre un meilleur contrôle sur leur composition où la présence des différentes classes de sons, ainsi que leur niveau sonore, sont définis par l'utilisateur en fonction de ses besoins. Leur utilisation trouve un intérêt dans le cadre des études perceptives des ESU et a déjà été réalisée pour, par exemple, étudier l'agrément sonore à partir de scènes sonores recréées directement par les participants et estimer l'influence de la présence des sources sonores [Lafay *et al.*, 2014], ce que ne permet pas l'écoute faite en ville [Adams *et al.*, 2008, Liu *et al.*, 2014] ou en laboratoire à partir d'enregistrements audio [Guastavino *et al.*, 2005, Cain *et al.*, 2013]. Ces scènes sonores simulées peuvent également trouver un intérêt pour tester et développer des outils de traitement du signal [Komatsu *et al.*, 2016, Geiger et Helwani, 2015] où l'utilisation de scènes sonores simulées permet d'éviter l'étape d'annotation manuelle, qui peut être longue et fastidieuse.

Toutefois, si l'utilisation de scènes sonores simulées a de nombreux avantages, la question de leur validité écologique est importante. En effet, la réalisation de *soundwalks* ou l'utilisation d'enregistrements sonore permettent de baser ces études sur des scènes sonores réelles et donc de mieux s'assurer que les outils sont bien adaptés à ces environnements.

En conséquence, il y a un intérêt certain à savoir correctement simuler des environnements sonores urbains de façon à obtenir une complexité et un réalisme suffisant, c'est-à-dire assimilable à des enregistrements faits en ville. Cette tâche reste un défi qui n'a, pour l'instant, pas été relevé et n'est pas trivial puisque l'environnement sonore urbain est un milieu extrêmement variable à la fois temporellement (à un endroit donné, les sources sonores varient constamment) et spatialement (d'un quartier à un autre, les sources ne sont pas les mêmes). Ce chapitre décrit notre méthode de création des corpus de scènes sonores urbaines simulées. Dans une première partie, plusieurs approches permettant de simuler de tels environnements sont résumées, puis l'outil retenu est présenté en détail. Les étapes menant à la création d'une base de données, appelé corpus élémentaire, et de deux corpus d'évaluation sont ensuite exposées. On présente enfin le test perceptif et les résultats obtenus démontrant le réalisme d'un des corpus.

4.1 Création de scènes sonores : choix d'une méthode

Dans une première partie, deux approches pour réaliser des scènes sonores urbaines sont présentées : l'auralisation et la simulation de scènes sonores.

4.1.1 Auralisation d'ESU

Une des premières approches possibles pour restituer un ESU est d'utiliser les techniques d'auralisation [Forssén *et al.*, 2009]. Cette méthode vise à modéliser l'évolution temporelle d'un signal sonore $M_i(t)$ en un point i en prenant en compte les différentes sources sonores $s_j(t)$ présentes ainsi que l'environnement spatial et les effets qu'ils génèrent sur la propagation des

sources sonores $\delta_{ij}(t)$. Cette méthode équivaut à modéliser les équations 1.1b et 1.1c. En choisissant le type et le nombre de sources et l'environnement urbain, il est alors envisageable de déterminer l'environnement $M_i(t)$. Pour cela, on réalise un produit de convolution entre la réponse impulsionnelle d'une rue, obtenue soit par sa mesure soit par sa modélisation par un logiciel (CATT-acoustics, I-Simpa ...), avec un signal sonore, enregistré dans des conditions d'anéchoïcité ou bien synthétisé. Cette étape correspond à l'équation 1.1b du chapitre 1. La restitution de l'ESU et son évolution dans le temps peuvent alors être écoutés [Vorländer, 2007]. Cette tâche reste toutefois complexe pour un tel environnement :

— La mesure de réponses impulsionnelles des rues [Picaut *et al.*, 2005] est une tâche complexe à réaliser puisqu'elle nécessite un dispositif expérimental conséquent qui doit être utilisé avec des conditions les plus neutres possibles (faible bruit de fond, conditions météorologiques neutres).

— La modélisation numérique des rues est alors la voie la plus souvent choisie car elle offre plus de possibilité, mais cela nécessite tout de même de simplifier l'environnement (allure des façades, présence de petits mobiliers urbain) afin de limiter les temps de calculs.

— Les effets de propagation du son en tenant en compte des phénomènes de diffusion, de réflexions dans un milieu urbain sont encore difficile à modéliser avec un rendu réaliste [Schissler *et al.*, 2014].

— La modélisation dynamique des sources sonores n'est faite que pour certaines sources sonores, comme le trafic routier ou ferroviaire, en utilisant des modèles dynamiques pour simuler leur déplacement. Ce sont alors parfois des enregistrements audio qui permettent de modéliser les autres sources sonores, ce qui permet de simplifier la modélisation mais restreint également le contrôle par l'utilisateur.

[Stienen et Vorländer, 2015] résument ces différents aspects, les questions soulevées et les champs d'applications que permet l'auralisation des environnements sonore urbains. Si cette tâche reste complexe, il existe tout de même quelques outils comme le logiciel *MithraSON* du CSTB qui propose de générer des ESU [1]. À partir d'un quartier modélisé, les sources sonores liées au trafic sont générés en temps réel à l'aide d'une synthèse granulaire et d'un modèle dynamique de trafic. L'ensemble des autres sources sonores (voix, oiseaux, cloche...) est basé sur des enregistrements audio qui sont ensuite intégrés à l'ESU. La propagation des signaux est générée à l'aide d'une méthode de tirs de rayons. Même si les résultats permettent une forte immersion, grâce à la spatialisation du son par l'écoute binaurale, cette méthode reste complexe à implémenter et nécessite des ressources numériques importantes.

4.1.2 Simulateur de scènes sonores

Une autre approche pour simuler l'environnement sonore urbain consiste à le considérer selon une combinaison additive d'évènements sonores (ou objets sonores), S_i, qui enrichi un bruit de fond (ou texture sonore), B [Nelken et De Cheveigné, 2013] :

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

$$M(n) = \sum_{i=1}^{N} S_i(n) + B \tag{4.1}$$

où n est un indice temporel. La catégorie bruit de fond inclut des sons longs (plusieurs minutes) dont les propriétés acoustiques ne varient pas (ou très peu) dans le temps comme le bruit généré par un trafic continu, le chant des oiseaux dans un parc ou les voies des enfants dans une cours de récréation. La catégorie *évènement* équivaut à des sons brefs (de 1 à plusieurs secondes), répartis dans le temps, qui émergent du bruit de fond pour être perçu individuellement par un auditeur (un voiture qui passe, des bruits de pas, une sonnerie de téléphone...).

Les évènements sonores S_i ayant la même origine sonore sont regroupés dans une même classe de son. Par exemple, on appelle *voiture*, la classe de son qui résume l'ensemble des sons qui sont générés par des voitures. Le défi est alors de disposer d'un nombre suffisant de classes de sons qui elles-mêmes regroupent suffisamment d'échantillons audio variés pour pouvoir recréer la diversité de l'ESU. Plusieurs outils ont été développés dans le but d'étudier la perception du paysage sonore (ou *soundscape* en anglais) et l'influence de la présence des différentes source sonores [Valle *et al.*, 2009, Finney et Janer, 2010]. Le simulateur TAPESTREA [Misra *et al.*, 2007] se base sur l'extraction de signaux sonores issus d'enregistrements, la classification de ces signaux (sinusoïdal, transitoire ou bruit de fond) et leur modulation afin de les insérer dans des mixtures sonores. Les fichiers audio modifiés peuvent alors être placés dans une scène sonore soit de manière bouclée, c'est-à-dire qu'un évènement sonore sera placé N fois dans un intervalle de temps, soit plus précisément en situant temporellement son emplacement. Ces techniques présentent l'avantage de s'appuyer sur des sons réels issus directement d'enregistrements sonores, et non des sons synthétisés ainsi que de permettre la modification des sons extraits selon de nombreux paramètres ainsi que d'avoir une grande maitrise dans la construction des scènes sonores. La limite de cette technique est la phase d'extraction où les évènements sonores doivent soit avoir un rapport *signal/bruit* élevé, soit ne pas présenter de recouvrement temporel et fréquentiel avec d'autres sources sonores. Sans cela, l'extraction des signaux est moins performante. De plus, dans le but d'obtenir des scènes sonores urbaines, il faut veiller à ne pas trop modifier les objets sonores afin d'éviter l'apparition d'artefacts qui réduiraient leur réalisme. D'autre simulateurs se basent sur des bases de données de sons pré-existantes comme chez [Bruce *et al.*, 2009] et [Rossignol *et al.*, 2015].

Dans l'outil de Bruce et Davies [Bruce *et al.*, 2009], l'utilisateur a la possibilité de choisir les sources sonores dans la scène, d'ajuster le niveau sonore et de choisir la position de la source. Leur base de données de sons est issue d'enregistrements audio réalisés par leur soin basé sur un nombre de sources défini selon des précédentes interviews et des marches sonores réalisées. Leur simulateur fut ensuite utilisé dans le cas de la synthèse de scènes sonores urbaines afin de déterminer les classes de sons les plus influentes [Davies *et al.*, 2014]. Enfin le simulateur développé par [Rossignol *et al.*, 2015], *SimScene*, propose à l'utilisateur de gérer un ensemble de paramètres (classe de son, position des évènements, émergence des évènements sonores par rapport au bruit de fond...) modélisés par des valeurs moyennes complétées par des écarts-types.

Cette particularité permet à l'utilisateur soit de définir précisément la position des évènements sonores (« je veux un sifflement d'oiseaux toutes les 5 secondes »), soit de générer des variations aléatoires que *SimScene* gère (« je veux un sifflement d'oiseau toutes les 5 ($\pm$ 2) secondes »). En plus de ces spécificités, cet outil a déjà été utilisé pour des études relatifs au payage sonore [Lafay *et al.*, 2015] où l'outil permet facilement la création de scènes sonores. Également, le simulateur a permis la réalisation de corpus de jeu de donnée pour le DCASE challenge [Stowell *et al.*, 2015] dans le cas de la tâche de détection d'évènements sonores [Lagrange *et al.*, 2015]. En raison de son fonctionnement et de son utilisation déjà éprouvée pour des ESU, le simulateur *SimScene* a été choisi pour réaliser les corpus de scènes sonores urbaines.

4.2 Présentation de *SimScene*

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

— le rapport *évènement/bruit de fond* (abrégé EBR pour *Event Background Ratio*),

— le temps de présence moyen d'une classe de son,

— l'occurrence moyenne d'une classe de son dans une scène,

— l'intervalle temporel τ entre chaque audio d'une même classe de son,

— la présence d'un *fade in* et d'un *fade out* pour chaque échantillon.

Chaque paramètre est également complété par un écart-type qui instaure de la variabilité d'une scène à l'autre. Les sons sont ensuite sélectionnés aléatoirement dans la base de données et positionnés dans la mixture sonore, calibrés selon l'EBR renseigné. En plus d'un audio pour la mixture sonore globale, un audio pour chaque classe de son présent dans la scène est généré permettant de connaitre sa contribution exacte. Ici, ce sont toutes les classes de sons relatifs au trafic routier qui permettent d'estimer son niveau sonore exact, $L_{eq,tr.}$ dans la scène.

En parallèle, *SimScene* génère 3 fichiers images (l'évolution temporelle du niveau sonore, le spectrogramme et un *piano Roll* pour visualiser la répartition dans le fichier de chacune des classes, Figure 4.1), un fichier texte résumant les temps de présence de l'ensemble des sons présents dans la scène et un fichier .mat où se trouve la totalité des résultats et des paramètres de la scène.

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

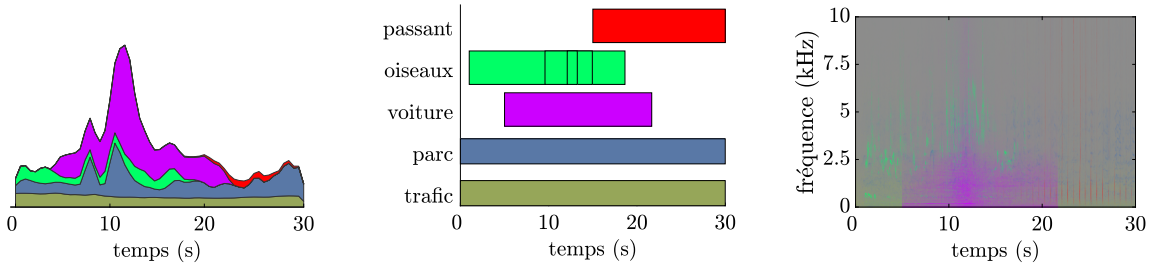


FIGURE 4.1 – Représentation temporelle (à gauche), *Piano Roll* (au centre) et spectrogramme (à droite) générés par *SimScene* d'une scène composée, d'un bruit de fond *trafic* (en vert foncé) et *parc* (en gris) et d'évènements *oiseaux* (en vert), *voiture* (en magenta) et *passant* (en rouge).

La génération de scènes avec *SimScene* peut se faire selon 2 modes. Dans le mode *abstract*, l'utilisateur renseigne lui-même les échantillons sonores présents dans la scène et chaque paramètre permettant de créer des scènes complètement artificiels. Dans le mode *replicate*, le schéma de la scène s'appuie sur un fichier texte où la position des évènements sonores (début et fin) et leurs classes de sons correspondantes sont détaillées. Ce mode permet de reproduire des scènes avec la même organisation temporelle que des enregistrements audio qui ont été annotés.

Pour offrir des scènes audio de qualité suffisante, *SimScene* nécessite de posséder une base de données de sons isolés, appelée corpus élémentaire, devant être suffisamment représentatif des sons entendu dans un ESU. De plus, la qualité de chaque audio (rapport Signal/Bruit, échantillonnage) doit être suffisante pour que leurs juxtapositions ne viennent pas détériorer le rendu final. Pour ces travaux, un corpus élémentaire de sons isolés, dédié à l'environnement urbain, a été créé.

4.3 Création d'un corpus élémentaire d'échantillons audio

4.3.1 Recherche en ligne des échantillons audio

La base de données de sons utilisée dans cette étude comprend un ensemble de classes de sons isolés (oiseaux, voiture, klaxon...) qui contiennent chacune plusieurs échantillons (*oiseaux01.wav*, *oiseaux02.wav*...) pour permettre une grande variabilité dans les mixtures sonores créées. La plupart des échantillons sont trouvés sur des sites en ligne de sons [3] [4] et à l'aide de la base de données constituée dans [Salamon *et al.*, 2014]. Leur base de données comprend en tout plus de 8000 fichiers audio, collectés également sur le site *freesound.org*, d'une durée inférieure à 4 secondes, répartis en 10 classes de sons : ventilation, klaxon de voiture, enfants qui joue, chien qui aboie, sonnerie, moteur en fonctionnement, coup de feu, marteau-piqueur, sirène et musique dans la rue. L'ensemble des échantillons a été trié afin de ne conserver que les audio ayant un rapport signal à bruit élevé et un échantillonnage de 44,1 kHz. À partir de la liste des noms des

3. `www.freesound.org`
4. `www.universalsoundbank.com`

fichiers originaux fournis avec cette base de données, les fichiers audio sont récupérés dans leur intégralité sur le site internet et intégrés dans la base de données.
Afin d'obtenir un rapport signal à bruit acceptable, certains audio ont été filtrés à l'aide du logiciel Audacity. D'autres signaux ont, quant à eux, été tronqués ou bien divisés en plusieurs fichiers afin d'en obtenir des durées convenables.

4.3.2 Enregistrements de passages de véhicules

S'il est possible de trouver l'ensemble des classes de son avec une qualité suffisante en ligne, dans le cas de la classe *voiture*, étant la source sonore d'intérêt, il était nécessaire de réaliser des enregistrements de passages de véhicules contrôlés sur une piste d'essai afin de posséder un ensemble varié et maitrisé de vitesses et de modèles de véhicules. Pour cela, 2 véhicules à motorisation essence (Renault Mégane, Renault Sénic) et 2 autres à motorisation diesel (Renault Clio, Dacia Sandero) ont été enregistrés en suivant un plan de mesures défini comprenant plusieurs vitesses stabilisées à différents rapports de vitesses ainsi que des phases d'accélération et de freinage du véhicule (voir Tableau 4.1).

TABLEAU 4.1 – Ensemble de mesures réalisées sur pistes avec des passages de véhicules à vitesses stabilisée (à gauche) et en accélération et freinage (à droite).

	Rapport	**1**	**2**	**3**	**4**	**5**
Vitesse stabilisée (km/h)	**20**	×	-	-	-	-
	30	-	×	×	-	-
	40	-	×	×	×	-
	50	-	-	×	×	-
	60	-	-	-	×	×
	70	-	-	-	×	×
	80	-	-	-	-	×
	90	-	-	-	-	×

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

Les enregistrements ont été réalisés sur la piste d'essais de l'Ifsttar de Nantes le 7 et 8 juillet 2016 à l'aide du système d'acquisition Sound Device 702 et d'un microphone omnidirectionnel. Sa position du microphone a respecté la norme de mesure de bruit au passage S 31-119 et fut donc situé à 7 m de la piste et à une hauteur de 1m50. Enfin, les conditions météorologiques étaient satisfaisantes (temps clair et dégagé, température à l'ombre de 25°C , vitesse moyenne du vent inférieure à 2 m/s). Les enregistrements sont ensuite extraits en fichiers audio en format .wav échantillonnés à 44,1 kHz.

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

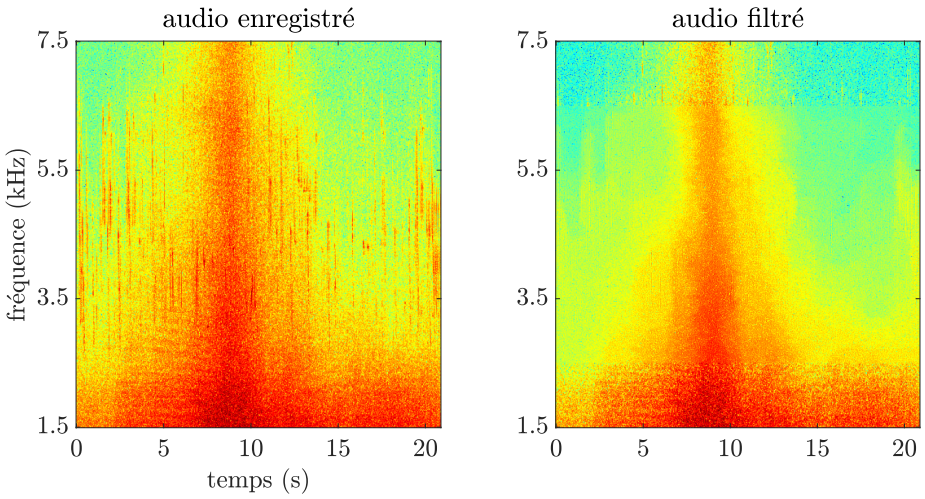


FIGURE 4.2 – Zoom du spectrogramme (nombre de points $w = 2^{12}$ avec 50 % de recouvrement) dans la bande de fréquence $[1500 - 7500]$ Hz d'un enregistrement de passage de véhicule (véhicule Renault, rapport 3, 40 km/h). À gauche, l'enregistrement original, à droite l'enregistrement filtré par le filtre médian.

de ce filtre est présenté en Figure 4.2. Même si elle reste persistante sur certains enregistrements, la présence des oiseaux est fortement atténuée sans toutefois dégrader la qualité perceptive du signal global du véhicule.

4.3.3 Composition du corpus élémentaire complet

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

Les enregistrements des passages de voitures sont, quant à eux, séparés en deux parties : les enregistrements issus des deux véhicules Renault Mégane et Renault Clio sont inclus dans le corpus élémentaire afin de construire les scènes sonores, les autres échantillons des deux autres voitures (Renault Scénic, Dacia Sandero) serviront dans les chapitres 5 et 6 afin de construire le dictionnaire de la NMF et ainsi éviter toute problématique de surapprentissage. Les échantillons sont ensuite séparés en deux classes de sons : *Voiture Ville* (si la vitesse stabilisée ou finale est inférieure ou égale à 50 km/h) et *Voiture Route* (si la vitesse stabilisée ou finale est supérieure à 50 km/h). L'ensemble des fichiers audio est en format .wav échantillonnés à 44,1 kHz. La base de données finale est résumée dans le Tableau 4.2 pour les évènements sonores et dans le Tableau 4.3 pour les bruits de fond sonores.

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

Tableau 4.2 – Composition de la base de données pour les évènements sonores.

Classe de son	Nombre	Classe de son	Nombre
Aboiement de chien	34	Porte de maison	8
Avion	11	Porte de voiture	5
Balais	6	Roulement de valise	5
Bruit de chantier (marteau, perceuse ...)	12	Sirène	9
Bruit de rue	24	Sonnette	5
Camion	4	Toussotement	7
Cloches d'églises	8	Train	7
Klaxon	24	Tram	4
Oiseaux	30	Voiture à l'arrêt	7
Orage	3	Voiture ville	28
Pas dans la ville	11	Voiture route	16
Pas dans un parc	16	Voix (rire, 1 ou 2 mots)	24
Total	**308**		

Tableau 4.3 – Composition de la base de données pour les bruits de fond.

Classe de son	Nombre	Classe de son	Nombre
Brouhaha de foule	15	Oiseaux	25
Brouhaha parc	25	Pluie	14
Chantier	28	Trafic routier	9
Cours de récréation	12	Vent dans les arbres	15
Fontaine	9	Ventilation	10
Total	**162**		

source sonore n'a pas pu être déterminée. De la même façon, les sons relatifs à un chantier en construction (marteau-piqueur, marteau, perceuse) sont regroupés en une seule classe par soucis de simplification. À partir de ce corpus constitué, disponible en ligne [5], il est possible de réaliser des corpus de scènes sonores urbaines. En vue d'étudier le comportement de la NMF puis d'en évaluer les performances, deux corpus de scènes sonores urbaines sont construits.

4.4 Corpus d'évaluation *Ambiance*

Dans un premier temps le choix est fait de générer un corpus où la présence de chaque source est définie selon sa classe de son et où les niveaux sonores du trafic sont calibrés. Ce corpus a vocation à être utilisé pour étudier le comportement de la NMF selon certaines sources sonores isolées et selon la prédominance du trafic routier dans les scènes. Les étapes impliquées dans la construction de ce corpus sont présentées dans la Figure 4.3. Nommé *Ambiance*, ce premier corpus consiste en un ensemble de 6 sous-corpus de 25 scènes M ayant chacune une durée de 30 secondes. Chaque sous-corpus, la mixture sonore M_i mélange une composante *trafic* ($S_{tr.}$) avec

5. https://doi.org/10.5281/zenodo.1213793

une classe de son spécifique (appelée classe *interférante*) ($S_{int.}$), tel que,

$$M_i = S_{tr.,i} + S_{int.,i}. \tag{4.2}$$

La composante $S_{tr.}$ inclut les évènements sonores appartenant aux classes de sons *Voiture Ville* et *Voiture route*, qui correspondent aux passages des voitures, et les bruits de fond *Trafic Routier*. Le reste est résumé dans la composante interférante $S_{int.}$.

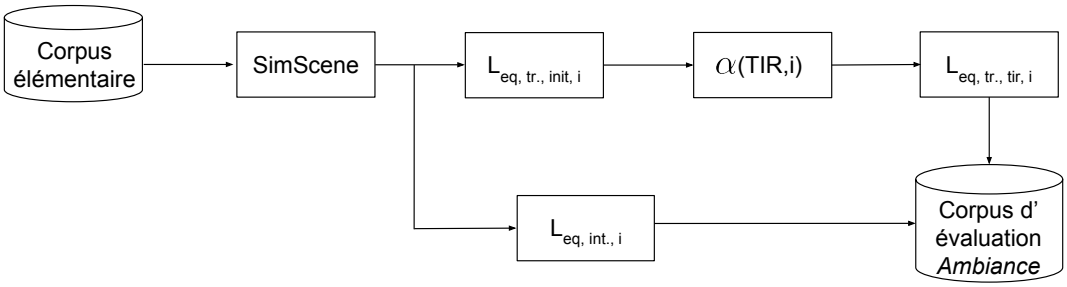


FIGURE 4.3 – Schéma bloc de la pondération du signal trafic selon la scène i et le TIR.

TABLEAU 4.4 – Résumé des classes de sons inclues dans les classes interférantes, seules les classes *alerte* et *transport* ne contiennent pas de bruit de fond.

Classe interférante	**Évènement**	**Bruit de fond**
alerte	- Klaxon - Sirène	-
animaux	- Oiseaux - Aboiement de chien	- Oiseaux
climat	- Orage	- Vent dans les arbres - Pluie
humain	- Voix - Bruit de pas	- Brouhaha de foule
transport	- Train - Tramway - Avion	-
mécanique	- Bruit de rue - Bruit de chantier	- Ventilation - Bruit de chantier

Ces 6 sous-corpus sont résumés dans le Tableau 4.4 avec les classes de sons inclues qui forment les classes interférantes. Chaque scène comprend un bruit de fond *trafic* ainsi que jusqu'à 5 passages de voiture. Pour les classes interférantes, leur présence dans chaque scène est systématique. Elle est définie selon un tirage d'une loi uniforme : une valeur aléatoire est tirée, selon sa valeur elle définit la présence ou non de la classe de son. Par exemple, dans le cas de la classe interférante *animaux*, qui comprend 2 classes de sons, il y a 33 % de chance d'avoir la classe *oiseaux*, 33 % de chance d'avoir des aboiements et 33 % de chance d'avoir les deux classes présentes. Pour le cas des signaux *alerte*, comme la durée d'un klaxon est plus brève que celle

d'une sirène, la répartition de la distribution est modifiée afin de mieux équilibrer leur présence temporelle (10 % de chance d'avoir une sirène, 80 % de chance d'avoir un coup de klaxon et 10 % d'avoir les deux dans la même scène). Dans le cas de *climat* et *mécanique*, d'autre bruits de fond peuvent également être présent là aussi équitablement répartit. Enfin pour la classe *humain*, la présence d'une foule en bruit de fond est présente une scène sur deux. On résume dans la Figure 4.4 les spectres moyens sur les 25 scènes du signal des classes *trafic* et *interférant*.

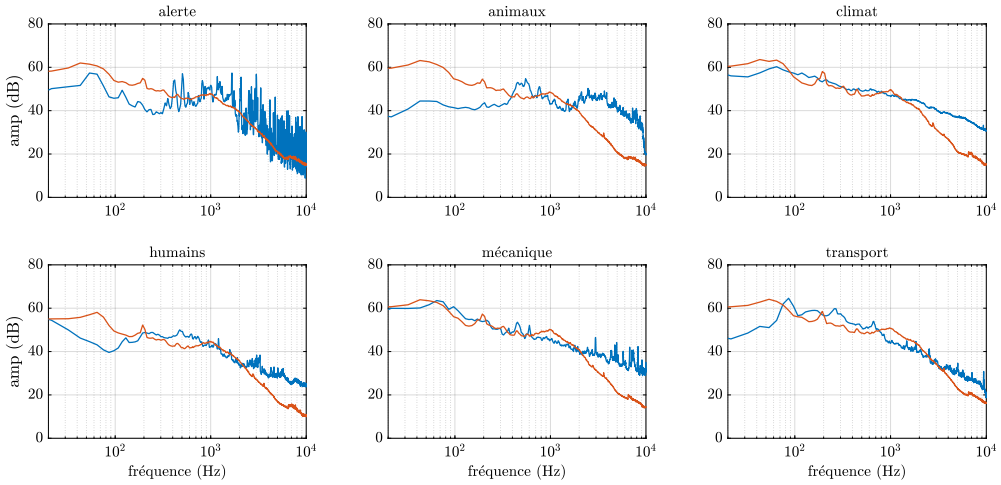


FIGURE 4.4 – Spectres sonores moyens des classes *interférante* (courbes en bleu) et *trafic* (courbes en rouge) pour chaque sous-corpus.

Dans le cas des sous-corpus *alerte* et *animaux*, les allures des spectres sont différentes de celles du *trafic* : ce sont des sons plus aigus et harmoniques dont les variations temporelles sont bien spécifiques. Le sifflement d'un oiseau ou le retentissement d'un klaxon sont plus ponctuels que le passage d'un voiture. À l'inverse, pour les autres sous-corpus, cette distinction est plus complexe puisque des composantes dans les basses-fréquences sont également présentes.

Chaque scène i générée possède alors un niveau sonore trafic initial global $L_{eq,tr.,init.,i}$ et un niveau sonore *interférant*, $L_{eq,int.,i}$. Elles sont ensuite, chacune, dupliquées 5 fois où le niveau sonore du trafic y est calibré par rapport au niveau sonore du signal interférant tel que, pour une scène i,

$$TIR = L_{eq,tr.,tir,i} - L_{eq,int.,i} \tag{4.3}$$

avec TIR le Rapport des niveaux sonores du trafic et de la classe interférante (*Traffic Interfering Ratio* en anglais) où $TIR \in$ {-12, -6, 0, 6, 12} dB. Ce TIR s'assimile au rapport *source-interférence* défini dans [Vincent *et al.*, 2006]. Pour cela, les fichiers audio relatifs au trafic sont pondérés par un coefficient α afin d'obtenir le niveau sonore souhaité selon le TIR avec

$$\alpha(TIR, i) = 10^{(TIR_{init,i} - TIR)/20} \tag{4.4}$$

où $TIR_{init.,i} = L_{eq,tr.,init.,i} - L_{eq,int.,i}$. Les étapes sont résumées sous la forme d'un schéma bloc dans la Figure 4.3. Lorsque $TIR < 0$ dB, le signal trafic est plus faible que le signal interférant, à l'inverse lorsque $TIR > 0$, le trafic devient la classe sonore prépondérante. La Figure 4.5 présente un exemple d'une scène sonore *alerte* pour 3 valeurs du TIR.

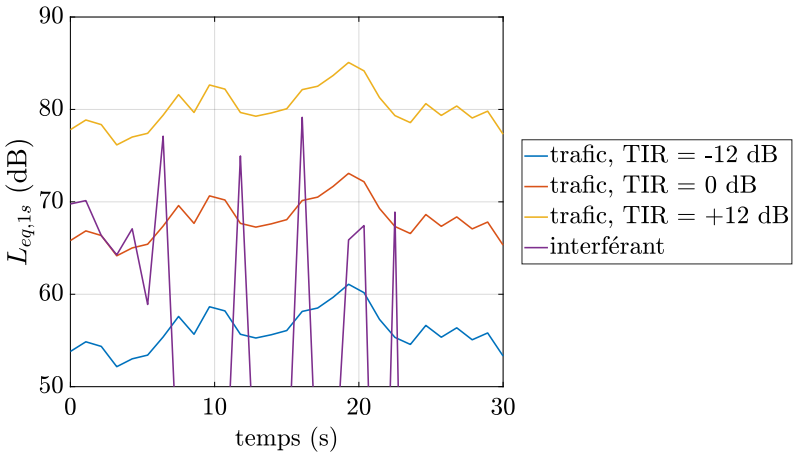


FIGURE 4.5 – Exemple de l'évolution du niveau sonore équivalent 1 seconde $L_{eq,1s}$ d'une mixture sonore extraite du sous-corpus *alerte* avec la composante *trafic* calibrée à $TIR \in \{-12, 0, 12\}$ dB.

En tout 750 scènes sont ainsi disponibles (6 sous-corpus × 25 scènes × 5 TIR) pour une durée totale du corpus de 6h30. Les scènes de ce corpus ne peuvent pas être assimilables à des enregistrements sonores réalisés en ville, mais permettront d'étudier le comportement des estimateurs dans le chapitre 5. Ce corpus de sons est disponible en ligne [6].

4.5 Corpus d'évaluation de scènes sonores urbaines réalistes

Un second corpus est généré, basé sur des enregistrements sonores réalisés en ville. Ce corpus d'évaluation de Scènes sOnores Urbaines Réalistes (corpus *SOUR*) a pour vocation de tester les performances de la NMF sur des scènes similaires à des enregistrements sonore faits en ville. Pour cela, un corpus de référence constitué d'enregistrements audio est obtenu pour ensuite être écouté et annoté. Ces annotations permettent alors de reproduire ces enregistrements en scènes simulées (dit *répliquées*) qui forment alors le corpus d'évaluation *SOUR*. L'ensemble des étapes est résumé sous forme de bloc dans la Figure 4.6.

4.5.1 Présentation des enregistrements audio de références

Les enregistrements audio de références sont issus du projet GRAFIC [Aumond *et al.*, 2017a] et ont été recueillis à pied dans le 13^{e} arrondissement de la ville de Paris sur un parcours compre-

6. `https://doi.org/10.5281/zenodo.1145855`

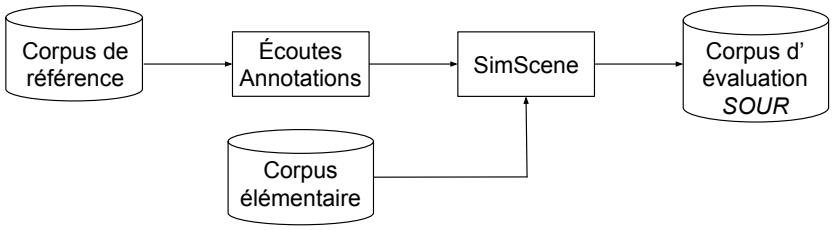


FIGURE 4.6 – Schéma bloc résumant la création du corpus d'évaluation de scènes sonores urbaines réalistes *SOUR*.

nant 19 points d'arrêts (Figure 4.7). Le parcours défini présente l'avantage de couvrir plusieurs ambiances sonores représentatives d'un environnement sonore urbain (Tableau 4.5).

FIGURE 4.7 – Parcours réalisé avec les 19 points de mesures avec le niveau sonore mesuré équivalent.

Ce trajet a été parcouru sur deux jours (le 23/05/2015, jour 1, et le 30/05/2015, jour 2), deux fois par jour (le matin puis l'après-midi) dans un sens (d'est en ouest, EW) et dans l'autre (d'ouest en est, WE). L'enregistrement est réalisé par un système d'acquisition équipé d'un microphone ASASense omnidirectionnel situé sur un sac à dos porté par l'opérateur [Aumond *et al.*, 2017a]. En tout, 76 enregistrements audio (19 points × 4 trajets) de 1 à 4 minutes sont disponibles.

4.5.2 Écoutes des scènes sonores

La première étape a consisté à réaliser une classification à l'écoute (Tableau 4.6), selon quatre ambiances sonores (*Parc*, *Rue calme*, *Rue bruyante*, *Rue très bruyante* [Can et Gauvreau, 2015]), des enregistrements sonores à partir des indications fournies dans [Aumond *et al.*, 2017a] (résumé dans le Tableau 4.5).

Tableau 4.5 – Résumé des 19 points de mesures avec l'ambiance générale [Aumond *et al.*, 2017a].

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

Tableau 4.6 – Classification des scènes par ambiances sonores.

Jour	trajet	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
1	**EW**																			
1	**WE**																			
2	**EW**																			
2	**WE**																			

Parc | Rue calme | Rue bruyante | Rue très bruyante | Non renseigné

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

4.5.3 Annotation des enregistrements sonores

L'annotation des 74 enregistrements est ensuite réalisée consistant à écouter chaque fichier audio et à estimer la nature des sources sonores présentes ainsi que leur temps de présence. Pour chaque enregistrement, l'ensemble des annotations est résumé dans un fichier texte. Un exemple d'annotation est présenté dans le Tableau 4.7.

Tableau 4.7 – Exemple d'un fichier d'annotation pour la scène 1-EW-07.

évènements	$\mathbf{t_{init}}$ **(s)**	$\mathbf{t_{fin}}$ **(s)**
bruit rue	0,00	8,50
voix	0,00	44,00
camion	1,00	56,10
voix	36,50	42,30
voiture Ville	52,00	63,00
voix	59,00	66,50

De ces annotations, il est possible d'estimer, par ambiance sonore, les sources sonores qui caractérisent leur bruit de fond, les classes de sons des évènements sonores ainsi que leur densité (nombre d'évènement par minute). Ces informations peuvent être utilisées pour pouvoir créer des scènes par le mode *abstract* de *SimScene* (Tableau 4.8). Le niveau sonore moyen est une donnée qui est à considérer avec prudence. En effet, le fichier audio de calibration n'a pas été fourni avec les enregistrements. Il n'a pas été possible d'établir avec certitude les niveaux sonores exacts des scènes. Une calibration relative a toutefois été réalisée sur l'ensemble des enregistrements audio avec l'aide des niveaux sonores renseignés par la Figure 4.7 afin d'obtenir, par scène, des niveaux sonores équivalents.

Sur l'ensemble des scènes sonores, 11 classes de sons sont identifiées en tant qu'évènement sonore (*trafic routier, voix, sifflements d'oiseaux, bruit de rue, bruit de pas, porte de maison, porte de voiture, chantier, klaxon, sonnette, sirène*) et 3 classes de sons sont présentes en tant que bruit de fond sonore (*brouhaha de foule, sifflements d'oiseaux, trafic routier continu*). Les sources sonores les plus communes sont *voiture*, *voix* et *bruit rue*. En outre, en plus des classes de sons résumées dans le Tableau 4.8, de nombreuses autres classes de sons (*aboiement de chien*, *bruit de balais*, *toussotement*, *passage d'avion*, *roulement de valise*) entendues interviennent plus sporadiquement (nombre d'évènement/min $< 0{,}1$) et sont susceptibles d'être présentes dans les quatre ambiances sonores. La composition des environnements sonores diffère naturellement selon les différentes ambiances sonores : dans *Parc* la voix et les oiseaux sont les bruits de fond sonores principaux permettant d'établir l'ambiance sonore adéquate, les sons les plus émergeant sont alors ceux relié à la présence humaine et aux oiseaux. Puis, plus les ambiances sonores sont dominées par la classe *trafic*, moins les émergences sonores sont élevées en raison du fort niveau du bruit de fond sonore *trafic*. Les classes de sons « naturel » (*oiseaux)* disparaissent alors progressivement au profit de celles liés aux activités humaines. Dans le cas de l'ambiance sonore *Rue calme*, les émergences sont plus élevées en raison d'un bruit de fond plus faible. Notons que dans *Rue calme*, *bruyante* et dans *Parc*, le décompte des voitures est assez aisé. Il l'est beaucoup moins dans *rue très bruyante* où un flot de véhicules peut être présent, le comptage y est alors très délicat car les véhicules peuvent être considérés à la fois comme bruit de fond et évènements sonore. Sans étude perceptive sur le débit de véhicules à partir duquel les passages de

TABLEAU 4.8 – Niveau sonore et description des classes de sons les plus récurrentes dans l'environnement urbain (nombre d'évènements sonore par minute > 0,1/min).

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

véhicules deviennent un flux, une moyenne de 1 véhicule par seconde est alors considérée comme raisonnable. Un contrôle à l'écoute permet de vérifier que le rendu est satisfaisant. Le rapport nombre d'évènement/min renseigné dans le Tableau 4.8 est donc soumis à une forte incertitude mais reste cependant cohérent avec les indications du débit moyen fournis dans [Aumond *et al.*, 2017a] ($\approx$ 2000 véhicules/heure).

4.5.4 Reproduction des enregistrements audio

Afin d'obtenir des scènes les plus réalistes que possibles, le choix a été fait de reproduire les 74 enregistrements à l'aide de leur annotation et du mode *replicate* de *SimScene*. Ce choix permet ainsi de s'assurer que la disposition des évènements sonores dans les mixtures sonores est la plus proche d'une structure temporelle écologiquement valide. Les durées cumulées par ambiance sonore sont résumées dans la Tableau 4.9. Avec un nombre de scènes plus importante, *rue calme* est naturellement l'ambiance dont la durée cumulée est la plus longue, l'ambiance *parc* étant alors la plus courte.

TABLEAU 4.9 – Durées cumulées par ambiance du corpus *SOUR*.

ambiance sonore	**N**	**durée (s)**
Parc	8	960
Rue calme	35	4636
Rue bruyante	23	3366
Rue très bruyante	8	1285
total	74	10 247

La difficulté dans la génération des scènes sonores réside surtout dans l'estimation du *event background ratio* pour les évènements sonores qui doit être cohérent par rapport à l'ambiance souhaitée. La détermination de sa valeur et de la variance correspondante s'est donc faite progressivement à l'écoute afin d'obtenir un rendu satisfaisant. Pour vérifier que la répartition des sons entre les éléments *trafic* et *interférant* dans chaque scène reste cohérent par rapport à l'ambiance sonore qui lui est assignée, le TIR dans chaque scène est calculé et résumé en Figure 4.8. La valeur du TIR moyen augmente linéairement avec l'ambiance sonore entre -9 dB et 17 dB. L'évolution du TIR à travers les ambiances traduit correctement la présence de plus en plus forte du trafic. Par ailleurs, la plupart des scènes possèdent un TIR positif et donc une part du trafic plus importante que celle de la classe interférante. Par rapport au corpus *Ambiance*, ce corpus privilégie donc plus des valeurs du TIR positifs.

Les scènes sonores sont ensuite calibrées non pas selon les niveaux sonores des enregistrements qui ont servi à les construire puisque leurs niveaux sonores exacts ne sont pas connus mais selon le niveau sonore moyen par ambiance sonore, résumé dans le Tableau 4.8. Cette étape n'influe en rien sur la suite de l'étude car les scènes sont construites d'un point de vue

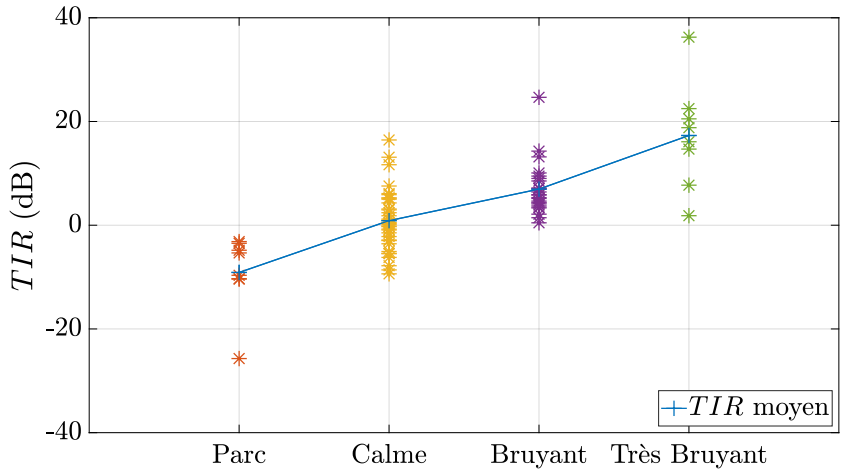


FIGURE 4.8 – Valeurs du TIR par scène et moyennés par ambiance sonore pour le corpus *SOUR*.

relatif, c'est-à-dire que les scènes sonores et l'ambiance auxquelles elles appartiennent ne sont pas liées à leur niveau sonore global mais aux différentes classes de sons présentes et à leurs émergences par rapport au bruit de fond. En vue de déterminer le niveau de bruit du trafic, le facteur déterminant est alors la contribution de cette source sonore dans la scène et non le niveau sonore global de celle-ci. Cette calibration a surtout pour objectif d'homogénéiser les différentes scènes sonores dans chaque ambiance. Ce corpus de sons est également disponible en ligne [7]

Dans la suite du document, les scènes issues du mode *replicate* de *SimScene* seront appelées « scènes répliquées » en raison du processus de duplication. Les scènes originelles sont, quant à elles, nommées « scènes enregistrées ». L'ensemble des ces scènes répliquées forment le *corpus d'évaluation SOUR*. L'annexe E résume la correspondance entre le nom des scènes enregistrées, nommées en fonction du jour, du sens du trajet et du point d'enregistrement (par exemple 1-EW-01), et le nom des scènes répliquées, nommées en fonction de l'ambiance sonore auxquelles elles appartiennent et d'un numéro d'identification (par exemple *Parc-01*).

4.6 Validation du réalisme du corpus d'évaluation *SOUR* par un test perceptif

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

7. https://doi.org/10.5281/zenodo.1184443

4.6.1 Mise en place du test

Ce test consiste à faire écouter à un panel d'auditeurs un ensemble de scènes sonores comprenant autant d'enregistrements sonores que de scènes reconstituées. Pour chaque scène, l'auditeur doit alors évaluer, sur une échelle de Likert à 7 points allant de « très peu réaliste » à « extrêmement réaliste », le réalisme de la scène qu'il vient d'entendre. L'hypothèse que nous souhaitons confirmer est que l'ensemble des scènes répliquées sont perçues de façon similaire aux scènes réalistes. Sur l'ensemble des 148 scènes (74 enregistrées, 74 répliquées), un ensemble de 40 scènes sont testés. Cet ensemble est composé dans une première moitié de scènes enregistrées choisies aléatoirement parmi les 74 enregistrements tout en prenant soin d'avoir une répartition équitable entre les ambiances sonores afin d'avoir suffisamment de diversité sonore. On extrait alors 5 scènes issues de l'ambiance *Parc*, 6 issues de *Rue calme*, 4 de *Rue bruyante* et 5 de *Rue très bruyante*. Pour chaque audio, 30 secondes sont ensuite sélectionnés aléatoirement. La seconde moitié du corpus est alors composée des mêmes 30 secondes des scènes répliquées respectives. Si l'hypothèse est vérifiée, nous supposerons que si le réalisme de ces 20 scènes répliquées est perçu de la même manière que les 20 scènes enregistrées, celui-ci pourra être étendu aux 54 autres scènes répliquées puisqu'elles sont construites sur le même processus. Un récapitulatif des fichiers audio sélectionnés et de la position des 30 secondes extraites sont résumés dans le Tableau 4.10.

Tableau 4.10 – Résumé des 40 audio composant l'ensemble des scènes testées avec les temps d'extraction des 30 secondes d'audio, l'identifiant et le nom des fichiers audio originaux.

ambiance	$\mathbf{t_{deb}}$	$\mathbf{t_{fin}}$	**id**	**scènes enregistrées**	**id**	**scènes répliquées**
Parc	41,7	71,7	1	1-EW-03	21	replicate-1-EW-03
	20,5	50,5	2	1-EW-08	22	replicate-1-EW-08
	38,2	68,2	3	1-EW-10	23	replicate-1-EW-10
	56,2	86,2	4	2-EW-03	24	replicate-2-EW-03
	38,5	68,5	5	2-WE-19	25	replicate-2-WE-19
Rue calme	20,0	50,0	6	1-EW-05	26	replicate-1-EW-05
	135,5	165,5	7	1-WE-06	27	replicate-1-WE-06
	28,6	58,6	8	1-WE-14	28	replicate-1-WE-14
	38,6	68,6	9	2-EW-13	29	replicate-2-EW-13
	110,7	140,7	10	2-WE-10	30	replicate-2-WE-10
	109,3	139,3	11	2-WE-05	31	replicate-2-WE-05
Rue bruyante	19,8	49,8	12	1-EW-01	32	replicate-1-EW-01
	211,6	241,6	13	1-EW-18	33	replicate-1-EW-18
	8,8	38,8	14	2-EW-02	34	replicate-2-EW-02
	57,5	87,5	15	1-WE-02	35	replicate-1-WE-02
Rue très bruyante	69,9	99,9	16	1-EW-16	36	replicate-1-EW-16
	75,6	105,6	17	1-WE-16	37	replicate-1-WE-16
	34,6	64,6	18	2-EW-16	38	replicate-2-EW-16
	87,3	117,3	19	2-WE-15	39	replicate-2-WE-15
	87,1	117,1	20	2-WE-18	40	replicate-2-WE-18

Pour limiter les erreurs statistiques dues aux variations de concentration du sujet lorsque les tests sont trop longs, chaque auditeur écoute un sous-corpus de 20 audio ; la durée du test n'excède alors pas 10 minutes. Comme les auditeurs n'évaluent plus l'ensemble des scènes mais seulement une partie, il faut définir un plan d'écoute qui répartit équitablement l'ordre de succession des écoutes. Pour cela, on réalise un plan expérimental en « Bloc Équilibré Incomplet » (BEI) [Pagès et Périnel, 2007]. En analyse sensorielle, un BEI permet d'élaborer l'ordre d'évaluation des produits testés pour chaque panéliste en évitant que des biais statistiques apparaissent (effet de rang, du juge, de succession...). Il se construit à partir de plusieurs variables :

— le nombre de blocs J (appelé ici auditeur),

— le nombre de traitements à tester, B (qui correspondant au nombre total d'extraits sonores dans le test),

— le nombre de traitements testé par juge, K (qui équivaut au nombre d'écoutes réalisées par chaque auditeur),

— le nombre de répétitions d'un traitement, R (qui équivaut au nombre de fois qu'un extrait audio est écouté),

— le nombre de répétitions d'une paire de traitement, λ (qui estime le nombre de fois qu'un couple de produit est testé successivement durant le test).

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

$$B \geq K, \tag{4.5a}$$

$$JK = BR, \tag{4.5b}$$

$$\lambda = R\frac{K-1}{B-1} \tag{4.5c}$$

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

La dénomination « incomplète » provient du fait que les juges n'évaluent pas tous les produits testés (condition 4.5a). La dénomination « équilibré », quant à elle, provient du fait que chaque juge évalue un même nombre de produits (K), que ces produits sont évalués un même nombre de fois (R) et que toute paire de produits est évaluée un même nombre de fois (λ).

Plusieurs paramètres ont été choisis et justifiés au début de la partie : le nombre d'extraits sonores testé a été établi à 40 ($B = 40$) pour un nombre d'extraits audio évalué par auditeur fixé à 20, ($K = 20$). La principale difficulté reste à obtenir la participation de J personnes pour ce test. Ce nombre est alors fixé à $J = 50$ en cela que ce nombre est suffisant et facilement atteignable en un temps raisonnable. À partir des variables J, B et K, le nombre R de réplication est fixé à 25. Toutefois, ces valeurs impliquent que la condition 4.5c n'est pas validée ($\lambda = 9,69 \notin \mathbb{N}$) et donc que les contraintes que l'on s'impose ne permettent pas d'obtenir un plan équilibré. Deux solutions sont alors possibles : la première serait de modifier certains paramètres pour

trouver l'équilibre. Or le nombre d'auditeur, $J = 50$, parait un nombre maximal raisonnable à atteindre tout comme le nombre de fichiers audio à tester K. Avec ces 2 contraintes fixées, il n'est pas possible d'obtenir un plan d'écoute satisfaisant. La deuxième solution, qui semble alors la plus adaptée, est de réaliser un plan optimal [Pagès et Périnel, 2007]. Dans ce cas, pour une configuration $[J, K, B]$ donnée, un algorithme d'échange détermine un plan qui satisfait le plus possible son équilibre (sans toutefois l'atteindre parfaitement). En d'autres termes, cet algorithme permet de déterminer la suite de produits testés par chaque juge qui permettra de respecter au mieux les conditions 4.5. Le plan optimal X_{opt} en fonction des conditions J, K et R est réalisé sous le logiciel R à l'aide la fonction *optimaldesign* fourni par le package *SensoMineR* [Lê et Husson, 2008].

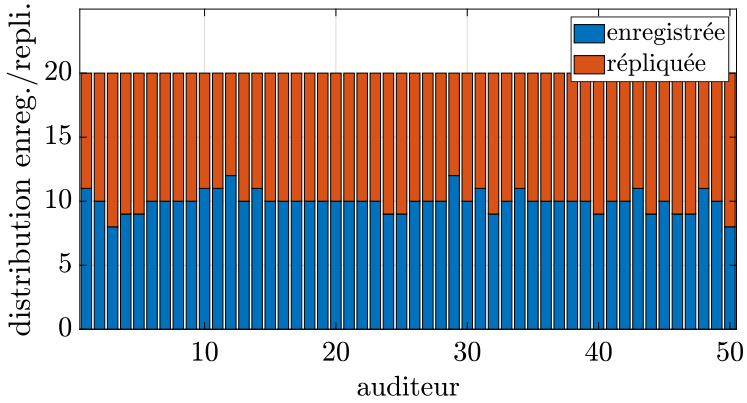


FIGURE 4.9 – Distribution des scènes audio pour chaque juge selon leur type (enregistré ou répliqué) : en bleu la quantité de scènes enregistrées évaluées par le juge et en rouge le nombre de scènes répliquées. La somme de ces deux parties équivaut au nombre de scènes testées K par juge.

Le plan obtenu correspond alors à l'ordre d'écoutes des scènes audio pour chaque juge. Avec ce plan, chaque juge écoute un mélange de scènes enregistrées et répliquées qui n'est pas nécessairement identique. Si la plupart des auditeurs écoutent un même nombre de scènes enregistrées et répliquées, d'autres sont susceptibles d'écouter plus de scènes d'un type que d'un autre. Au maximum, certain juges écoutent 8 scènes d'un type et 12 scènes d'un autre (Figure 4.9). De ce mélange, les auditeurs, durant le test d'écoute, n'évaluent donc pas nécessairement une scène enregistrée et sa version répliquée. L'optimisation du plan ne permet également pas d'avoir un nombre de réplications R constant sur l'ensemble des scènes testées mais variable évoluant dans l'intervalle $[20, 30]$ (Figure 4.10).

Une page web [8] est mis en ligne le 8 février 2017 permettant l'accès au test à un large public et s'est clôturé 12 jours plus tard. Chacun des 50 auditeurs écoute donc une succession de 20 extraits audio de 30 secondes dans un ordre établi par le plan optimal. Il leur est demandé de réaliser de test sur des enceintes ou un casque audio de qualité suffisante. Après avoir écouté

8. http ://soundthings.org/research/xpRealism

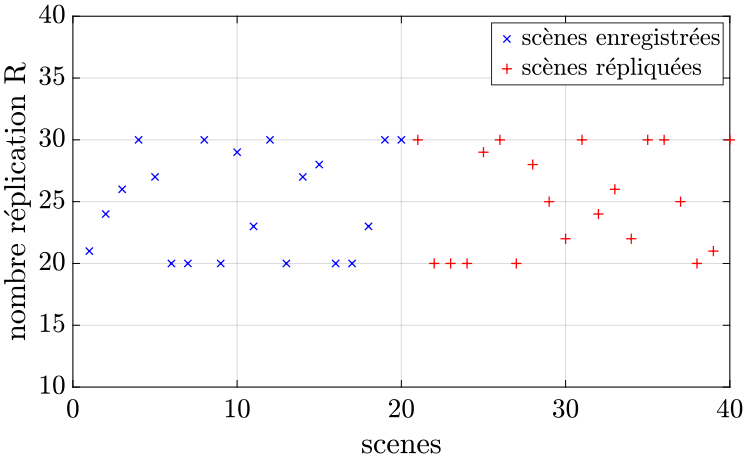


FIGURE 4.10 – Nombre de réplications, R, pour chaque scène obtenu dans X_{opt} avec comme combinaison $J = 50$, $B = 40$, $K = 20$. Les 20 premières scènes sont les scènes issues des enregistrements du projet GRAFIC, les 20 suivantes sont les scènes répliquées sous *SimScene*.

chaque extrait de 30 secondes, l'auditeur doit répondre à la question « La scène que vous venez d'entendre vous semble t-elle réaliste ? » en donnant une note entre 1 et 7. Chaque audio peut être réécouté autant de fois que voulu avant d'être évalué sans qu'il soit toutefois possible de revenir sur son évaluation. L'auditeur a également la possibilité de laisser un commentaire sur chaque audio pour justifier son choix. En fin de test, afin de connaitre le panel d'évaluateur, il leur est demandé de renseigner leur âge, leur sexe (H/F) et leur expérience quant à l'écoute de mixtures sonores urbaines.

Les fichiers résultats sont stockés également sous une page web [9] et téléchargeables sous le format .json pour ensuite être traités sous le logiciel Matlab.

4.6.2 Résultats

L'ensemble des résultats est soumis à différents tests statistiques afin de comparer les évaluation des scènes enregistrées et répliquées.

4.6.2.1 Constitution du panel

La Figure 4.11 résume, sous forme d'histogrammes, l'âge, le sexe et l'expérience des 50 auditeurs ayant participé au test. 2 personnes n'ont renseigné aucun de ces champs et une troisième personne a seulement omis de préciser son genre. Le panel est composé à 62 % d'hommes et à 32 % de femmes. La classe d'âge $[20, 30[$ est la plus représentée suivie de la classe $[30, 40[$ (26 %), $[50, 60[$ (18 %), $[40, 50[$ (10%) et enfin de la classe > 60 (4 %) . 62 % du panel a déclaré

9. http ://soundthings.org/research/xpRealism/responses/

n'avoir pas d'expérience dans l'écoute d'ambiances sonores urbaines.

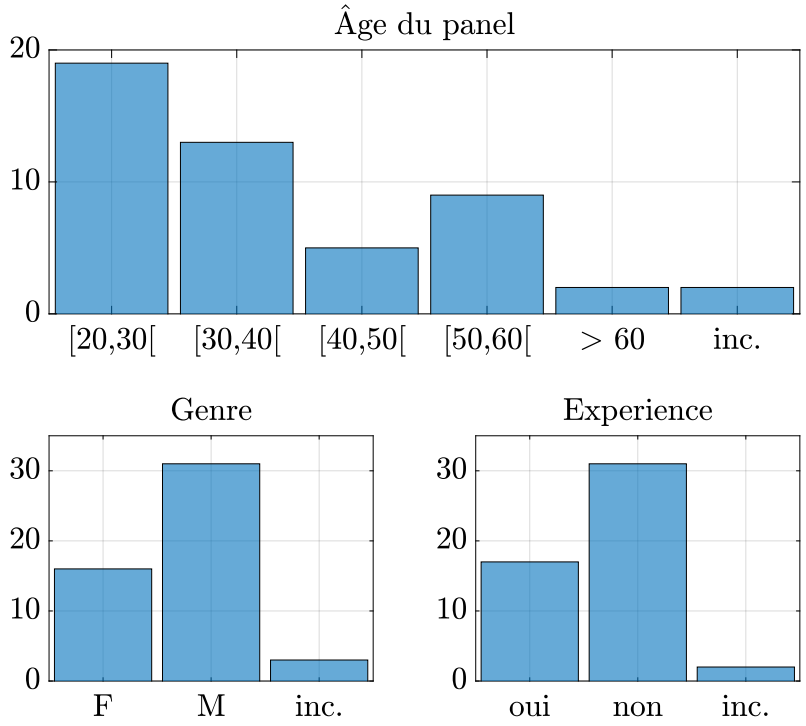


FIGURE 4.11 – Résumé des informations relatifs aux auditeurs.

4.6.2.2 Distribution des notes des scènes enregistrées et répliquées

Dans un premier temps, la distribution de toutes les notes de réalisme données par les auditeurs selon leur type (enregistrées et répliquées) est exprimée au travers d'un diagramme de type « boîte à moustache » (Figure 4.12). Cette représentation graphique permet de comparer plusieurs distributions en résumant pour chaque boîte la médiane (trait plein rouge), les valeurs du premier quartile au troisième quartile (boîte en bleue), la valeur maximale et minimale de la distribution (respectivement trait supérieur et inférieur en noir). À cela est également ajoutée la moyenne.

La répartition des notes pour les deux types de scènes est fortement similaire. Chaque type présente des valeurs identiques (médiane, valeurs extrêmes, quantiles). Seule la note moyenne permet de différencier les deux ensembles ($m_{En} = 4{,}93$ ($\pm$ 1,64) et $m_{Re} = 5{,}06$ ($\pm$ 1,56)) où la note moyenne des scènes répliquées est légèrement supérieure.

Afin d'étudier l'effet du type de scène sur le réalisme perçu, une analyse scène par scène est nécessaire (les différences de réalisme pouvant se compenser entre les scènes lors d'une analyse

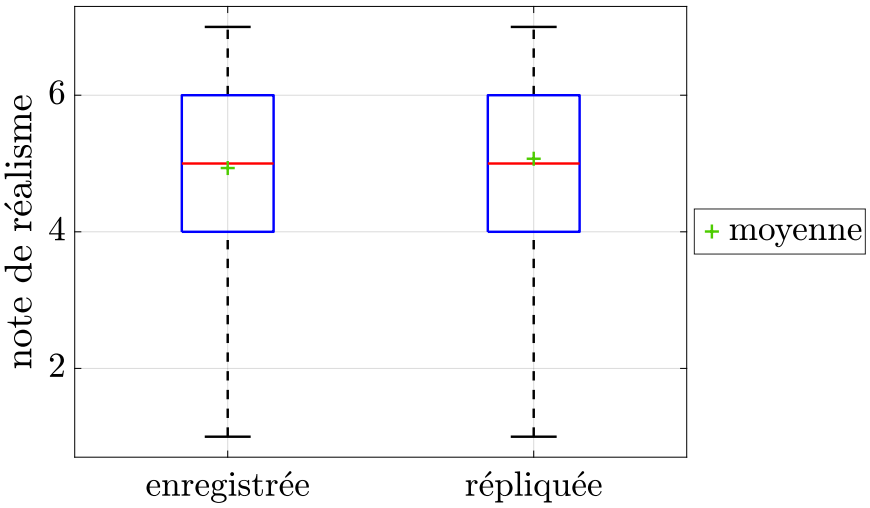


FIGURE 4.12 – Représentation en diagramme en boîte à moustache entre les scènes enregistrées et répliquées.

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

— les distributions des échantillons des deux catégories sont semblables (hypothèse *nulle* H_0),

— les deux distributions sont différentes, (hypothèse *alternative* H_1).

La statistique t est alors calculée :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \tag{4.6}$$

où $\bar{X}_i$, s_i et N_i sont, respectivement, la moyenne de l'échantillon, la variance et le nombre d'échantillons de la catégorie i ainsi que les degrés de liberté (DDL) du système :

$$DDL = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}}. \tag{4.7}$$

La statistique t avec le nombre de degrés de libertés (correction de Welch) sont alors utilisés

avec une loi de Student pour déterminer la valeur de la probabilité p (valeur p) qui permet de rejeter (ou non) l'hypothèse H_0 selon une valeur seuil de référence α (défini à 5 %) :

— si $p < \alpha$, les distributions considérées sont différentes, l'hypothèse H_0 est rejetée et H_1 est acceptée,

— si $p > \alpha$, l'hypothèse H_0 est acceptée, les distributions d'où sont issues les évaluations sont considérées identiques.

Par soucis de concision, seul l'ensemble des 20 valeurs p calculées et les boites à moustaches de chaque scène sont résumés dans les Tableaux 4.11 et 4.13.

Tableau 4.11 – DDL, valeurs t et valeurs p pour chaque test de Student mené entre les scènes enregistrées et répliquées ; en gras, les valeur p supérieures au seuil de signification de 5 %.

scène	1	2	3	4	5	6	7	8	9	10
DDL	48	51	51	51	51	57	57	60	60	60
valeur t	1,03	0,21	0,44	1,02	2,13	1,30	1,54	0,45	0,85	1,01
valeur p	**0,15**	**0,42**	**0,33**	**0,16**	0,02	**0,10**	**0,06**	**0,32**	**0,20**	**0,16**

scène	11	12	13	14	15	16	17	18	19	20
DDL	60	60	60	60	60	60	60	60	60	60
valeur t	0,29	0,39	1,15	0,30	0,84	0,32	0,82	2,20	0,24	0,45
valeur p	**0,38**	**0,35**	**0,13**	**0,38**	**0,20**	**0,37**	**0,21**	0,01	**0,40**	**0,33**

L'ensemble des tests de Student mené sur les 20 couples de scènes révèlent des valeurs p inférieures au seuil de signification α de 5 % pour seulement 2 scènes (scène 5 et 18). Ces résultats sont à mettre en parallèle avec la distribution des notes de ces scènes en Figure 4.13. Dans le cas de la scène 5, on observe que le réalisme moyen de la scène répliquée est plus important que celui de la scène enregistrée (Figure 4.13(a)). À l'écoute, la scène 5 issue de l'enregistrement est une scène calme avec peu d'évènements émergents et un bruit de fond, lui aussi, très calme. Il peut être supposé que sans identification claire d'évènements sonores, les auditeurs ont jugé le réalisme de la scène enregistrée moins correctement que sa version répliquée qui, elle, si elle possède aussi peu d'évènements, par son aspect simulé, parait toutefois plus identifiable. Enfin, pour la scène 18, les deux distributions sont toutes deux situées vers des notes élevées. Mais dans le cas de sa version enregistrée (Figure 4.13(b)), la distribution également plus large que la scène répliquée génère une valeur p sous le seuil de signification. Ainsi, malgré un test de Student qui validerait l'hypothèse H_1, on peut toute de même considérer que le réalisme des scènes répliquées 5 et 18 est satisfaisant au regard de la distribution de leur notes. Sur les 18 autres scènes , l'hypothèse H_0 n'est donc pas rejetée : le réalisme perçu des scènes testées du type *répliquée* n'est pas significativement différent de celui des scènes *enregistrées*. L'influence de l'évaluation des juges et l'influence de l'ambiance sonore sur le réalisme perçu sont observés, en complément, en Annexe B, à travers des analyses de variances (ANOVA).

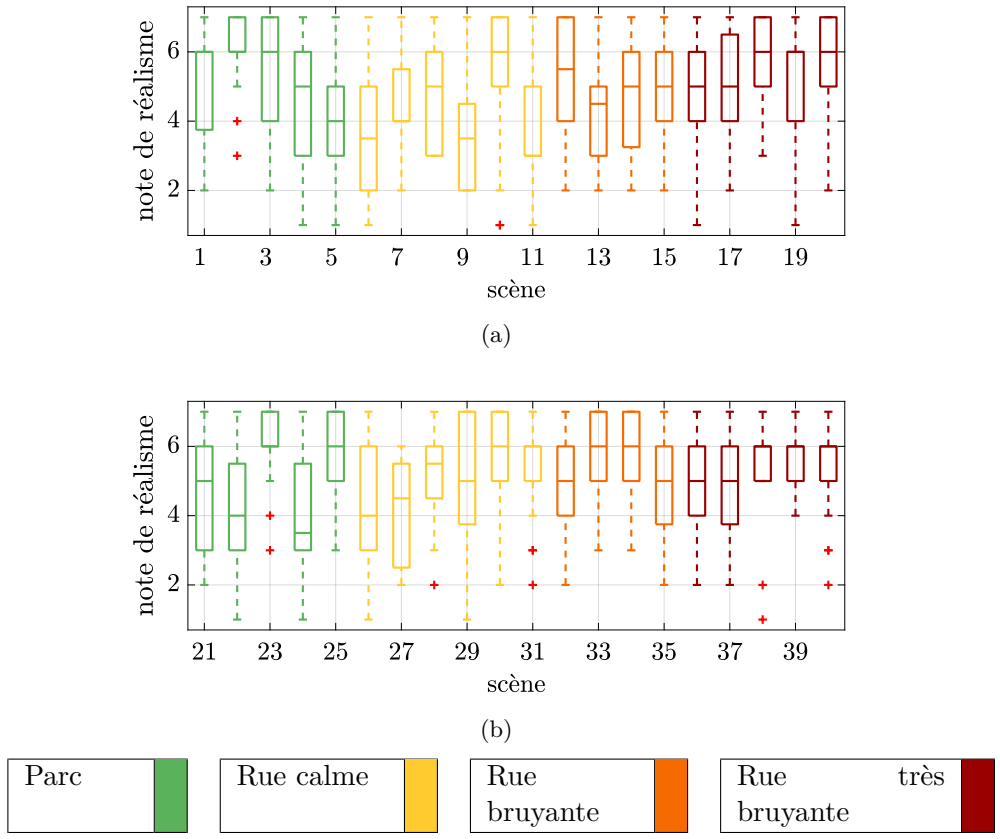


FIGURE 4.13 – Boites à moustaches pour les scènes enregistrées (a) et pour les scènes répliquées (b) classées selon leur ambiance sonore.

Enfin, il est intéressant d'étudier les commentaires laissés par les auditeurs sur plusieurs scènes. Ces derniers relèvent notamment des sons trop forts ou qui s'inscrivent mal dans les scènes (par exemple oiseaux trop forts dans la scène 21, bruits de pas trop fort dans la scène 24). Enfin certains extraits de voix ne sont pas suffisamment réalistes. En effet, lors de la phase d'écoutes des enregistrements, on remarque que les voix perçues, en dehors d'un brouhaha de foule, sont le plus souvent des bribes de conversations entre deux personnes ou au téléphone. Malheureusement, il n'a pas été possible de trouver des bases de données libres de conversations suffisamments réalistes pour être inclus dans les scènes. De nombreuses bases de données se concentrent, par exemple, sur la lecture de textes récités [El Ayadi *et al.*, 2011, Kominek et Black, 2004, Barker *et al.*, 2015] ce qui ne permet pas d'atteindre le réalisme souhaité. Les extraits de voix présents dans le corpus élémentaire sont des sons trouvés par défauts, brefs ("hello", "how are you ?") et qui diffèrent beaucoup des voix entendues en ville. Certains commentaires ont toutefois été laissé sur des scènes enregistrées : pour la scène 4, un auditeur trouve le sifflement des oiseaux trop fort « artificiel »ou dans la scène 8, un auditeur déclare que le trop grand nombre d'évènements sonores nuit au réalisme de la scène. Ces commentaires permettent alors de relativiser ceux laissés sur les scènes répliquées.

En conclusion, sur l'ensemble du corpus testé, il y a une similarité du réalisme perçu par

l'ensemble des participants entre les scènes enregistrées et répliquées. Même s'il y a des disparités entre les distributions des notes, l'évaluation du réalisme selon les types de scènes et les ambiances restent similaires (les notes restent comprises entre 4 et 6). Certains sons dans les scènes répliquées sembleraient toutefois mal calibrés, ce qui, sur certaines scènes, impactent leur note (la scène 24 par exemple, voir Figure 4.13(b)). Mais on retrouve également ce phénomène dans les scènes réelles, ce qui viendrait à relativiser cette mauvaise calibration. En conséquence, le réalisme perçu des scènes répliquées est considéré comme similaire à celles des scènes réalistes. De ces résultats, comme l'ensemble des scènes répliquées est réalisés avec le même dispositif, on généralise ce résultat à l'ensemble du corpus élémentaire *SOUR*.

4.7 Conclusion du chapitre

Le logiciel de simulation de scènes audio *SimScene* a été utilisé afin de constituer deux corpus de scènes sonores urbaines. Cette élaboration a nécessité la construction d'un corpus élémentaire qui contient des évènements sonores isolés (sifflement d'oiseaux, aboiement de chien, klaxon...) ainsi que des bruits de fonds (bruit de trafic routier continu, brouhaha de voix...). Puisque le trafic routier est la source sonore d'intérêt, des passages de voitures ont été enregistrés sur pistes afin d'obtenir des échantillons audio de qualité suffisante. De cette base de données, un premier corpus a été élaboré : le corpus d'évaluation *Ambiance*. Il comprend en tout 750 scènes de 30 secondes et se divise en 6 sous-corpus qui sont caractérisés par une classe de son générique (*alerte*, *animaux*, *climat*, *humains*, *transport* et *mécanique*. Chaque sous-corpus est composé de 25 scènes mixant un signal audio trafic avec un signal audio de la classe de son générique, appelée classe de son *interférante*. Chaque scène est alors dupliquée 5 fois où le niveau sonore du trafic est calibré selon celui de la classe de son interférante. Ce corpus a pour objectif de tester le comportement de la NMF suivant les classes de sons interférantes et des contributions du trafic différentes. Un second corpus a ensuite été élaboré, le corpus d'évaluation de Scènes Sonores Urbaines Réalistes *SOUR*, qui est la transcription d'enregistrements sonores urbains en scènes simulées. Afin d'estimer la qualité de ces scènes et notamment leur réalisme, une partie de ce corpus a été soumise à un test perceptif qui a révélé que les scènes simulées étaient perçues de façon similaires à des enregistrements audio. Les conclusions de ce test sont alors étendus à l'ensemble du corpus construit. L'intérêt de ce critère de réalisme est de pouvoir assimiler ces scènes à des enregistrements audio et ainsi d'estimer les erreurs que générerait la NMF sur de telles mixtures. Ce corpus a aussi pour vocation à servir l'ensemble des communautés scientifiques développant des outils de reconnaissance, de détection ou de séparation de sources dans un milieu urbain.

Chapitre 5

Étude du comportement de la NMF sur le corpus d'évaluation *ambiance*

Résumé

La NMF est appliquée sur le corpus d'évaluation *Ambiance* afin de visualiser le comportement de cette méthode sur de telles mixtures sonores. Les trois méthodes retenues (NMF supervisée, semi-supervisée et initialisée seuillée) sont testées pour de multiples configurations selon la composition du dictionnaire et de la β-divergence. Les résultats obtenus permettent de constater des performances variables des méthodes selon la prédominance du trafic. Lorsque celui-ci est faible, la NMF semi-supervisée se révèle l'approche la plus performante alors que la NMF supervisée génère de plus faibles erreurs quand le trafic est la source sonore principale. La NMF IS est alors la méthode qui offre le meilleur compromis dans son fonctionnement grâce à la mise à jour de son dictionnaire et à l'estimation du signal *trafic* par une méthode de seuillage dur.

Dans ce chapitre, on étudie le comportement des différentes versions de la NMF, présentées dans le chapitre 3, avec le premier corpus élémentaire *Ambiance*. Ce corpus présente l'intérêt d'être construit en mixant des classes de sons spécifiques dont la source sonore *trafic* est calibrée à différents niveaux sonores. Cet aspect permet d'étudier le fonctionnement des différentes versions de la NMF en fonction de la prédominance du trafic et de la nature des classes de sons et de déterminer les approches les plus efficaces et les plus adaptées à ces environnements. Dans un premier temps un rappel du corpus, des méthodes choisies et une présentation de la méthode de référence (ou *baseline* en anglais) sont exposés. Puis les étapes menant à la réalisation du dictionnaire et l'ensemble des facteurs expérimentaux sont détaillées. Enfin les résultats des calculs menés sont présentés et discutés.

5.1 Rappel de la méthode employée

Les étapes impliquées dans l'estimation du niveau sonore du trafic à partir de scènes sonores simulées sont d'abord rappelées. La Figure 5.1 résume la démarche générale.

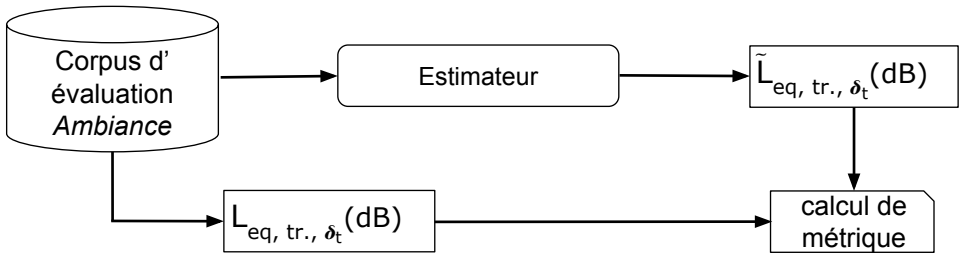


FIGURE 5.1 – Schéma-bloc de l'estimation du niveau sonore du bruit de trafic.

Le corpus d'évaluation *Ambiance*, présenté dans 4.4, est composé de scènes sonores classées en 6 sous-corpus : *alerte*, *animaux*, *climat*, *humain*, *transport*, *mécanique* (abrégés respectivement *al.*, *an.*, *cl.*, *hu.*, *tr. me.*). Chaque sous-corpus est lui-même divisé en 5 sous-ensembles qui comprennent, chacun, les mêmes 25 mixtures sonores M_i de 30 secondes, comprenant une classe de son *trafic* (qui inclut le bruit de fond routier ainsi que les évènements sonores *passages de voitures*), $S_{tr.}$, et une classe de son *interférante* (qui regroupe tous les autres sources sonores), $S_{int.}$:

$$M_i = S_{tr.,i} + S_{int.,i}. \tag{5.1}$$

La différence dans chacun des 5 sous-ensembles réside dans la calibration des niveaux sonores du signal *trafic*, $L_{eq,tr.}$, par rapport aux niveaux sonores de la classe de son *interférante*, $L_{eq,int.}$ tel que :

$$TIR = L_{eq,tr.} - L_{eq,int.} \tag{5.2}$$

avec 5 valeurs $TIR \in \{-12,\ -6,\ 0,\ 6,\ 12\}$ dB. Ce corpus permet de tester les performances de la NMF et son comportement face à différentes sources sonores avec une prédominance variable du trafic routier. En tout, le corpus est composé de 750 scènes (6 sous-corpus $\times$ 5 TIR $\times$ 25 scènes) pour une durée totale de 6h30.

Pour chaque TIR et chaque sous-corpus, les 25 scènes i sont soumises à un estimateur qui détermine le niveau sonore *trafic* de l'intégralité de la scène (temps d'intégration $\delta_t = 30$ secondes) en dB, $\tilde{L}_{eq,tr.,30s,i}$. Les 25 niveaux sonores sont ensuite comparés aux niveaux sonores exacts respectifs, $L_{eq,tr.,30s,i}$ à travers le calcul de métrique MAE (équation 1.15 définie dans la partie 1.5). Il est ensuite possible de déterminer les performances d'un estimateur sur l'ensemble des 6 sous-corpus pour chaque TIR en calculant sa moyenne,

$$MAE_{TIR} = \frac{\sum_{i=1}^{6} MAE_i}{6}, \tag{5.3}$$

ainsi que l'erreur globale sur l'intégralité du corpus *Ambiance* :

$$MAE_g = \frac{\sum_{i=1}^{6}\sum_{j=1}^{5} MAE_{i,j}}{6 \times 5}. \tag{5.4}$$

L'erreur MAE_g traduit l'erreur moyenne de l'estimateur faite sur l'intégralité du corpus d'évaluation et donc celle qui serait faite si aucune connaissance *a priori* sur l'environnement sonore n'était disponible.

5.2 Estimateur de référence

Dans un premier temps, un estimateur de référence (ou *baseline* en anglais) est nécessaire afin de comparer les performances de la NMF. La baseline choisie est un filtre passe-bas de fréquence de coupure f_c. L'hypothèse faite est que l'énergie située dans la bande passante est assimilable au signal *trafic*, $\tilde{L}_{eq,tr.}$. Le choix de cet outil est justifié par la présence de composantes basses-fréquences (principalement en dessous de 1000 Hz) dans les signaux trafic. Supprimer l'énergie au-delà parait donc une première approche envisageable. Cet estimateur consiste à représenter une mixture M_i sous la forme d'un spectrogramme puis à rejeter toutes les trames fréquentielles supérieures à f_c. La Figure 5.2 résume les étapes intervenant pour cet estimateur.

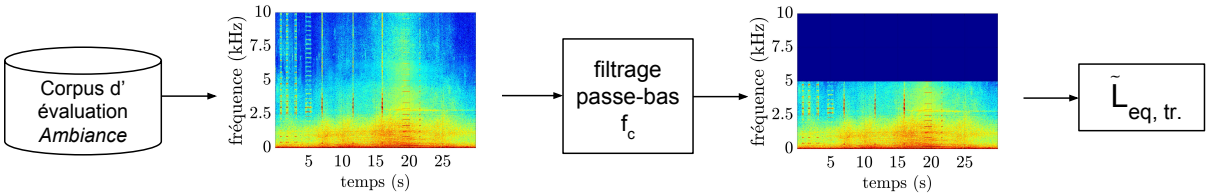


FIGURE 5.2 – Principe de l'estimateur *baseline* pour une mixture sonore filtrée à $f_c = 5$kHz.

Les fréquences de coupure choisies sont $f_c \in \{100, 500, 1k, 2k, 5k, 10k, 20k\}$ Hz. Le cas où $f_c = 20$ kHz correspond finalement au cas où aucun traitement du signal n'est réalisé sur les fichiers audio et où toutes les sources présentes sont assimilées au trafic, ce qui correspond à l'utilisation actuellement faite des mesures pour la correction de cartes de bruits issues de modèles prédictifs.

5.3 Estimateur basé sur la NMF

Le second estimateur est celui basé sur la NMF, présentée dans le chapitre 3. Pour rappel, cette méthode consiste à approximer le spectrogramme en amplitude $\mathbf{V}$ d'un signal audio par le produit de deux matrices, $\mathbf{W}$, un dictionnaire de spectres sonores, et $\mathbf{H}$, une matrice d'activation temporelle. La Figure 5.3 rappelle les différentes étapes présentes dans cet estimateur.

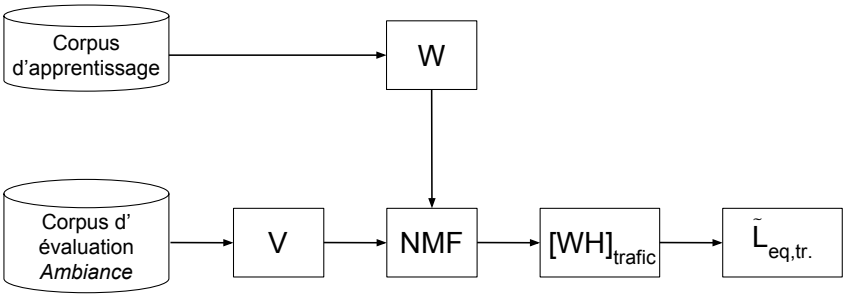


FIGURE 5.3 – Schéma bloc de l'estimateur NMF sur le corpus d'évaluation *Ambiance*.

5.3.1 Constitution du dictionnaire

Dans un premier temps, le dictionnaire **W** est construit, à partir d'un corpus d'apprentissage composé de 53 enregistrements audio des passages des voitures Renault Scénic et Dacia Sandero. Ces enregistrements ont été réalisés sur la piste d'essais de l'Ifsttar dans les mêmes conditions d'enregistrements que les 2 véhicules composants le corpus élémentaire de *SimScene* (voir partie 4.3.2). Ces 53 échantillons audio issus des enregistrements ne sont pas ceux utilisés dans la création des scènes sonores afin d'éviter tout problème de sur-apprentissage.

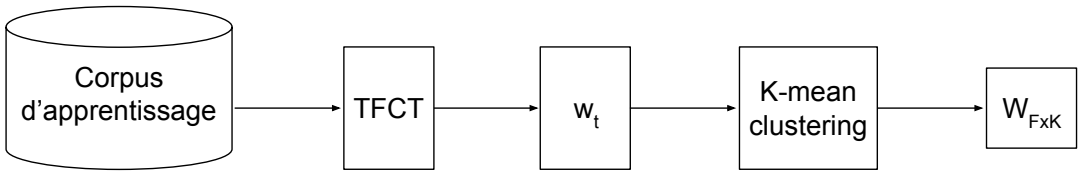


FIGURE 5.4 – Schéma bloc de la création du dictionnaire.

La constitution du dictionnaire est réalisée en trois étapes, résumées dans le schéma bloc en Figure 5.4 :

— chaque fichier audio est représenté au travers d'un spectrogramme, obtenu par une Transformée de Fourier à Court Terme (abrégé TFCT) (nombre de points $w = 2^{12}$ avec 50 % de recouvrement). Cette première étape permet d'obtenir pour chaque échantillon audio, de durées différentes, le même nombre de points en fréquences.

— Chaque spectrogramme est ensuite découpé en plusieurs fenêtres temporelles de durée $w_t \in \{0.5,\ 1,\ 2\}$ seconde(s). Dans chacune des fenêtres, la valeur efficace *rms* sur chaque trame fréquentielle est calculée. Ce procédé, qui équivaut à un sous échantillonnage, a pour but d'obtenir différentes représentations des spectrogrammes initiaux avec une description du contenu spectral plus ou moins précise. Dans le cas où $w_t = 0,5$ s, les spectres sonores contiennent plus de détails que dans le cas où $w_t = 2$ s. Les étapes de ce processus sont résumées en Figure 5.5 sur un extrait sonore de 3 secondes.

— Enfin, l'opération précédente générant un grand nombre d'éléments (2218 pour $w_t = 0{,}5$ s, 505 pour $w_t = 2$ s), une quantification vectorielle, opérée grâce à un algorithme de clustering

K-means, est appliquée en vue de réduire ce nombre à $K = \{25,\ 50,\ 100,\ 200\}$ et d'éviter la présence d'informations redondantes. Les K centroïdes obtenus par cet algorithme sont alors les éléments qui composent le dictionnaire $\mathbf{W}$.

En plus de ces étapes, on ajoute un cas où la valeur *rms* est calculée sur l'ensemble des spectrogrammes ($w_t = all$). Des 53 fichiers audio du corpus d'apprentissage, 53 spectres sont générés. Cette opération permet de baser la construction du dictionnaire sur les enveloppes spectrales des fichiers audio *trafic* et moins sur une description fine des spectres. Ces 53 spectres sont également soumis à l'algorithme de clustering mais avec cette fois, $K_{w_t=all} \in \{25,\ 50\}$. L'ensemble de ces facteurs expérimentaux et leurs modalités sont résumés dans le Tableau 5.1. L'intérêt de ces paramètres est de multiplier le nombre de formats que peut prendre le dictionnaire pour ainsi estimer l'influence de la forme et du nombre de ces éléments dans l'estimation du niveau sonore du trafic selon les différentes NMF. Enfin, chaque élément du dictionnaire, pour toutes les combinaisons testées, est normalisé selon la norme ℓ_1 telle que :

$$\|\mathbf{w}\|_1 = \sum_{f=1}^{F} w_f = 1. \tag{5.5}$$

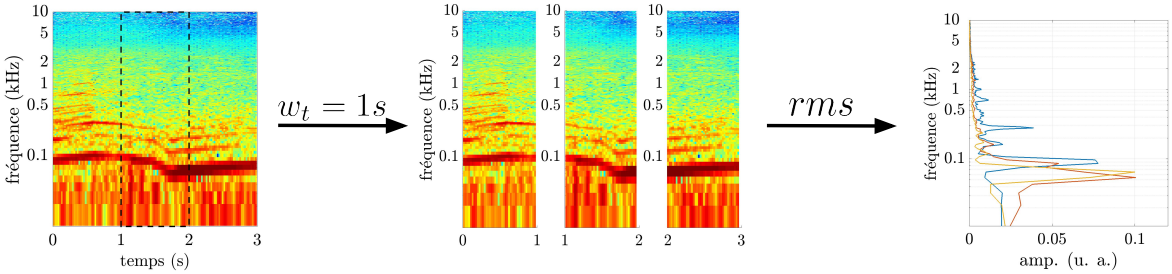


FIGURE 5.5 – Création des éléments de $\mathbf{W}$ sur un extrait de 3 secondes du passage d'une voiture pour une trame temporelle $w_t = 1$ seconde. À gauche, le spectrogramme du signal audio avec, en pointillés, une fenêtre de découpe. Au centre, le signal découpé en trois trames et dont les valeurs *rms* sont ensuite calculées générant 3 spectres.

5.3.2 Réalisation de la NMF

Chaque version du dictionnaire $\mathbf{W}$ est utilisée par l'estimateur du niveau de trafic. Cet estimateur est lui-même basé sur plusieurs versions de la NMF décrites précédemment (voir chapitre 3) : la NMF supervisée (NMF SUP), semi-supervisée (NMF SEM) et initialisée seuillée (NMF IS). Pour chaque NMF, 3 β-divergences sont utilisées : la distance Euclidienne ($\beta = 2$) (voir partie 3.3.1), la divergence de Kullback-Leibler ($\beta = 1$) (partie 3.3.2) et la divergence d'Itakura-Saito ($\beta = 0$) (partie 3.3.3). La NMF SUP dépend seulement des différentes versions du dictionnaire apprises et des valeurs de β. Le signal trafic est estimé selon l'équation 3.41 à partir de la matrice $\mathbf{H}$ obtenue. Dans le cas de la NMF SEM, le dictionnaire appris compose la partie fixe $\mathbf{W_s}$. Le nombre d'éléments du dictionnaire libre $\mathbf{W_r}$ est alors fixé à 2 ($J = 2$). Le signal *trafic* est ensuite déterminé par le relation 3.45. Enfin pour la NMF IS, les dictionnaires

appris correspondent aux dictionnaires initiaux $\mathbf{W_0}$ qui seront ensuite mis à jour. À chaque itération, les éléments du dictionnaire sont également tous normalisés selon la norme ℓ_1. Les dictionnaires obtenus $\mathbf{W'}$ sont ensuite soumis à l'étape d'extraction par seuillage qui implique également plusieurs facteurs expérimentaux :

— la représentation de la distance entre les dictionnaires initiaux et finaux $D_\theta(\mathbf{W_0}\|\mathbf{W'})$ (linéaire ou bien exprimé au travers d'une fonction sigmoïde ($\lambda = 2$)),

— le type de seuillage appliqué (dur ou *firm*),

— les valeurs des différents seuils respectifs (t_h pour le seuillage dur et $t_{f,1/2}$, les deux valeurs seuils pour le seuillage *firm*). Des études préliminaires ont permis de réduire la plage des valeurs de ces seuils à $t_h \in [0,30\ 0,70]$, $t_{f,1} \in [0,20\ 0,55]$ et $t_{f,2} \in [0,35,\ 0,70]$, chacun étant défini avec un pas de 0,01. Rappelons que pour le seuillage *firm*, $t_{f,1} \leq t_{f,2}$.

Malgré la réduction du nombre d'éléments par l'algorithme K-means, la taille des matrices $\mathbf{W}$ et $\mathbf{V}$ reste importante en raison du nombre de trames fréquentielles ($F = 2049$). En conséquence, ces deux matrices sont exprimées en bandes de tiers d'octave ce qui réduit les dimensions des matrices ($F_{1/3} = 29$). L'allure d'un spectre du passage d'une voiture en bandes fines et en tiers d'octave est représenté en Figure 5.6. La manipulation des matrices est alors plus rapide qu'avec les bandes fines et permet donc un gain en temps de calcul. Cette représentation a également d'autres intérêts :

— par son échelle logarithmique, elle décompose mieux les basses fréquences que les hautes fréquences ce qui permet de mieux focaliser la reconstruction du signal vers les bandes de fréquences d'intérêt.

— Cette représentation est également couramment utilisée dans le domaine de l'acoustique urbaine et environnementale, à la différence des MFCC. Notamment tous les réseaux de mesures en ville déterminent des valeurs du niveaux sonores en bandes de tiers d'octave. Cela en fait donc une représentation adaptée.

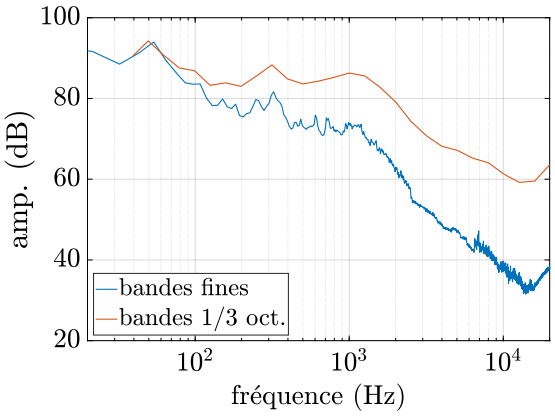


FIGURE 5.6 – Spectre en fréquence du passage d'une voiture en bandes fines (2049 points) et en bandes de tiers d'octave (29 bandes).

Enfin, le nombre d'itérations est fixé à 100 pour toutes les versions de la NMF étudiées. Également, si le niveau en sortie de l'estimateur est exprimé en dB, la NMF est réalisée avec le spectrogramme et le dictionnaire exprimés en grandeur linéaire (équivalent à la pression *rms* en Pa). Lorsque la composante *trafic* est estimée (par l'équation 3.41 pour la NMF supervisée, l'équation 3.45 pour la NMF semi-supervisée et l'équation 3.47 pour la NMF Initialisée Seuillée), la racine de la moyenne au carré (abrégé rms pour *root mean square*) est calculé pour chaque trame temporelle n :

$$p_{trafic,n} = \sqrt{\left(\frac{\sum_F \mathbf{v_{trafic,n}}^2}{1.5}\right)} \tag{5.6}$$

où 1,5 est un facteur correctif dû à l'effet de fenêtrage *hanning* dans le spectrogramme $\mathbf{V}$ [1]. Ce niveau sonore est ensuite exprimé en dB par l'équation 1.13 où il peut être exprimé selon la période d'intégration δ_t choisie (équation 1.14).

5.3.3 Résumé des facteurs expérimentaux

De nombreux facteurs expérimentaux sont donc présents dans cette expérience, chacun ayant différentes modalités. Le Tableau 5.1 résume l'ensemble de ces paramètres dont les différents sous-corpus, valeurs du TIR et facteurs expérimentaux liés aux estimateurs. La Figure 5.7 représente les différents facteurs expérimentaux impliqués dans l'estimateur ainsi que leurs liens de dépendances.

Tableau 5.1 – Facteurs expérimentaux et leurs modalités utilisés pour le corpus d'évaluation *Ambiance*.

facteurs expérimentaux	**modalités**							**nombre de modalités**
sous-corpus	alerte (al.)	animaux (an).	climat (cl.)	humain (hu.)	transport (tr.)	mécanique (me.)		6
TIR (dB)	-12	-6	0	6	12			5
estimateur	filtre passe bas	NMF SUP	NMF SEM	NMF IS				4
f_c (kHz)	0,1	0,5	1	2	5	10	20	7
w_t (s)	0,5	1	2	*all*				4
K	25	50	100	200				4
β	0	1	2					3
représentation	linéaire	sigmoïde						2
seuillage	dur	*firm*						2
seuil dur $\mathbf{t_h}$	de 0,30 à 0,60 avec un pas de 0.01							31
seuil firm $\mathbf{t_{f,1}}$	de 0,20 à 0,55 avec un pas de 0,01							36
seuil firm $\mathbf{t_{f,2}}$	de 0,35 à 0,70 avec un pas de 0,01							36

Pour l'estimateur filtre, c'est donc 210 combinaisons qui sont réalisées (6 sous-corpus × 5

1. voir `http://blog.prosig.com/2009/09/01/amplitude-and-energy-correction-a-brief-summary/`

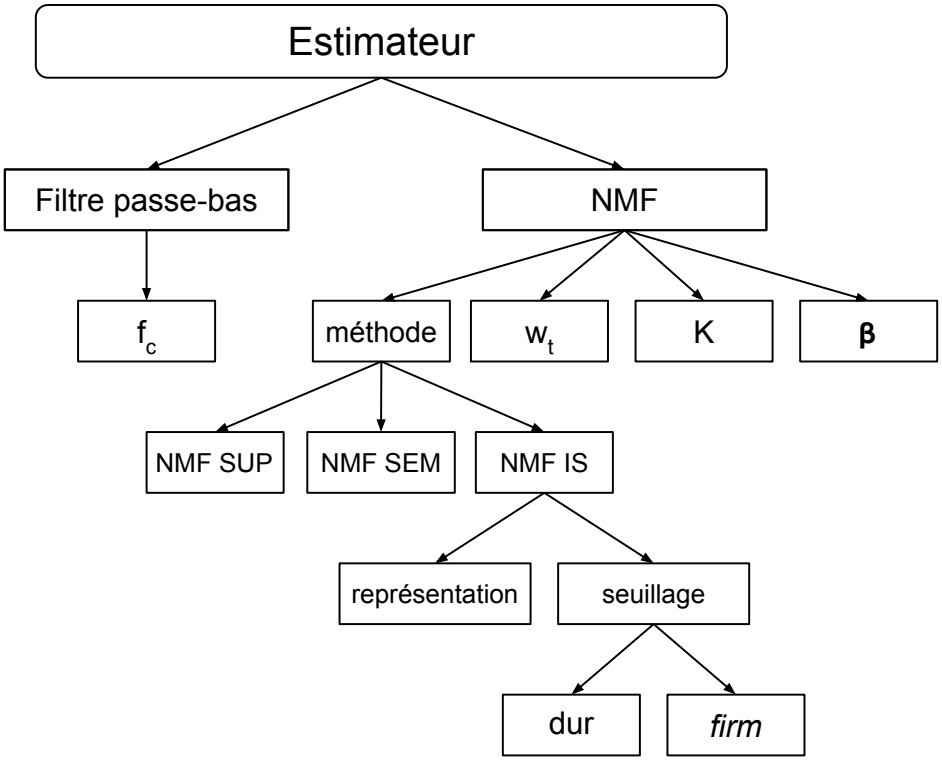


FIGURE 5.7 – Organigramme des différents facteurs expérimentaux impliqués dans l'outil *estimateur*.

TIR × 7 f_c). Dans le cas de la NMF SUP et SEM, ce sont respectivement 1260 combinaisons qui sont évaluées (6 sous-corpus × 5 TIR × (3 w_t × 4 K + 1 w_t × 2 K) × 3 β). Dans le cas de la NMF IS, ce nombre est beaucoup plus élevé (2 789 640) en raison des nombreuses valeurs seuils (6 sous-corpus × 5 TIR × (3 w_t × 4 K + 1 w_t × 2 K) × 3 β × 2 représentations × (31 + 1076)).

Les calculs exhaustifs de toutes les combinaisons expérimentales sont réalisés avec le logiciel Matlab à l'aide de l'outil expLanes [2] qui permet la réalisation d'expériences numériques, de gérer la distribution des facteurs expérimentaux et de leurs modalités et de collecter les nombreux résultats générés. Les programmes mis en place et développés sont disponibles en ligne [3].

5.4 Performances de l'estimateur *baseline*

Les résultats issus de l'estimateur *baseline* sont d'abord présentés. Dans un premier temps, les erreurs MAE_g (équation 5.4) générées par les estimations réalisées par chaque fréquence de coupure sont résumées dans le Tableau 5.2 ainsi que les erreurs MAE_{TIR}.

La plus faible erreur MAE_g sur l'intégralité du corpus est obtenue pour une fréquence de coupure f_c = 500 Hz (MAE_g = 2,89 (± 2,84) dB). À l'inverse, c'est naturellement pour la

2. `http://mathieulagrange.github.io/expLanes`
3. `https://github.com/jean-remyGloaguen/trafficNoiseEstimationNMF.git`

TABLEAU 5.2 – Erreurs MAE_g et MAE_{TIR} en dB de l'estimateur *baseline* selon f_c sur l'ensemble du corpus *Ambiance* et pour chaque TIR. En gras-rouge l'erreur MAE_g la plus faible, en gras-noir, les erreurs MAE_{TIR} les plus faibles selon les fréquences f_c.

f_c (Hz)	MAE_g	MAE_{-12}	MAE_{-6}	MAE_0	MAE_6	MAE_{12}
100	3,21 (± 1,06)	**4,55 (± 1,55)**	**2,92 (± 0,42)**	2,36 (± 0,71)	2,97 (± 0,41)	3,25 (± 0,32)
500	**2,89 (± 2,84)**	7,39 (± 3,00)	3,44 (± 1,65)	**1,17 (± 0,24)**	1,03 (± 0,26)	1,45 (± 0,13)
1000	3,36 (± 3,63)	9,44 (± 2,03)	4,78 (± 1,34)	1,62 (± 0,54)	**0,36 (± 0,10)**	0,61 (± 0,06)
2000	3,83 (± 4,01)	10,30 (± 1,57)	5,65 (± 1,05)	2,25 (± 0,49)	0,62 (± 0,15)	**0,11 (± 0,02)**
5000	4,51 (± 4,43)	11,95 (± 0,20)	6,70 (± 0,16)	2,82 (± 0,10)	0,87 (± 0,04)	0,20 (± 0,02)
10000	4,64 (± 4,51)	12,19 (± 0,08)	6,90 (± 0,07)	2,95 (± 0,05)	0,92 (± 0,02)	0,22 (± 0,01)
20000	4,69 (± 4,52)	12,25 (± 0,05)	6,96 (± 0,05)	3,00 (± 0,03)	0,97 (± 0,01)	0,26 (± 0,00)

fréquence de coupure de 20 kHz que l'erreur est la plus importante puisque l'intégralité des sources sonores est considérée ($MAE_g = 4{,}69$ (± 4,52) dB). En détaillant les erreurs MAE_{TIR}, on constate que la fréquence de coupure f_c correspondant à l'erreur minimale évolue en fonction de la prédominance du trafic : celle-ci augmente avec le TIR. Avec l'augmentation de la présence des voitures, il est nécessaire d'augmenter la fréquence f_c afin de conserver le plus possible l'énergie sonore de cette source. À l'opposé, lorsque le TIR est négatif, le filtre le plus performant sera celui qui est susceptible d'éliminer suffisamment d'énergie pour ne pas prendre en considération les sources sonores interférantes. Ce comportement est caractéristique d'un outil de détection qui peut réaliser des faux positifs (ce qui correspond à la prise en compte de la classe interférante dans le signal *trafic*) et rater des évènements (ce qui revient ici à supprimer trop d'énergie du signal *trafic*). Dans le cas du filtre $f_c = 500$ Hz, on détaille par sous-corpus et par TIR les distributions des erreurs relatives $\tilde{L}_{eq,tr.} - L_{eq,tr.}$ (Figure 5.8) sous la forme de diagramme de type « boite à moustache »et les erreurs MAE par des histogrammes (Figure 5.9).

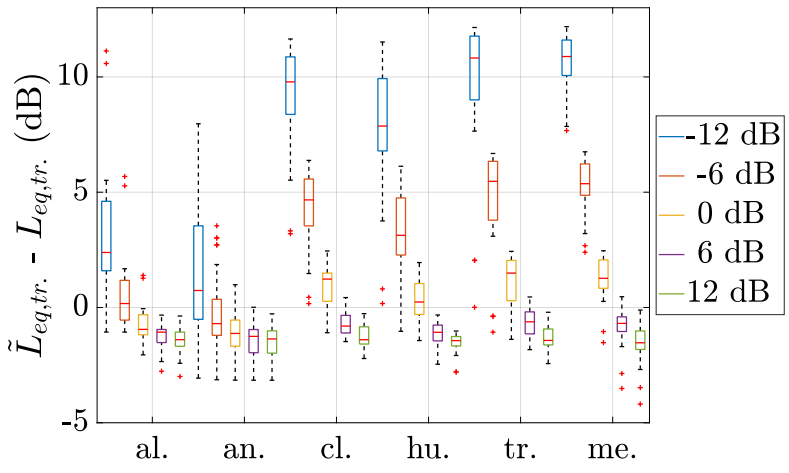


FIGURE 5.8 – Distributions des erreurs relatives entre les niveaux estimés et exactes pour chaque sous-corpus et chaque TIR pour l'estimateur filtre passe-bas à la fréquence de coupure $f_c = 500$ Hz.

La Figure 5.8 permet de visualiser correctement le comportement du filtre passe-bas selon

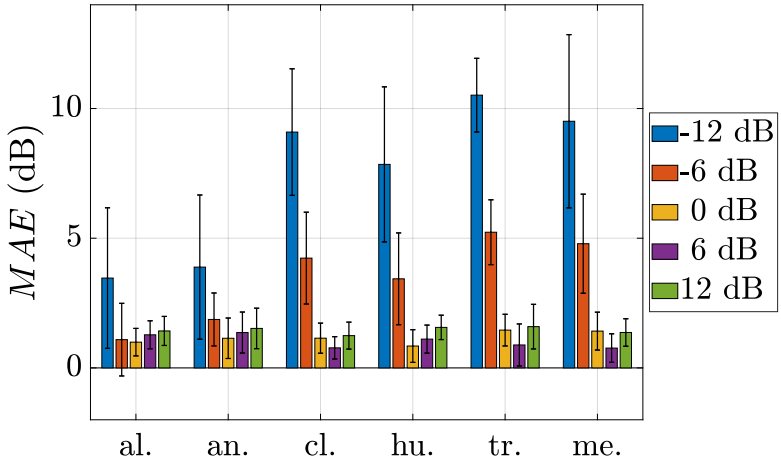


FIGURE 5.9 – Erreurs MAE pour chaque sous-corpus et chaque TIR pour l'estimateur filtre passe-bas à la fréquence de coupure $f_c = 500$ Hz.

l'évolution du TIR là où la Figure 5.9 permet de visualiser l'erreur MAE produite. Lorsque $\tilde{L}_{eq,tr.} - L_{eq,tr.} > 0$ dB, les estimations du niveau sonore trafic sont surestimées en raison de la prise en compte d'une partie des composantes de la classe *interférante*. Ce cas est alors présent notamment lorsque $TIR \in \{-12, -6\}$ dB. Cette surestimation est plus importante dans le cas des sous-corpus *climat*, *humain*, *transport* et *mécanique* qui possèdent plus de composantes basses-fréquences que les sous-corpus *alerte* et *animaux*. Les erreurs MAE pour ces classes de sons y sont alors plus faibles que pour les quatre autres classes de sons. Avec l'augmentation du TIR, la présence du trafic devient prédominante, l'estimateur basé sur le filtre passe-bas sous-estime alors le niveau exact du trafic en raison de la suppression d'une trop grande quantité d'énergie de cette composante. L'erreur MAE diminue toutefois à des valeurs plus faibles (MAE < 2 dB) avec systématiquement une erreur MAE plus élevée pour $TIR = 12$ dB. Dans le cas extrême où $f_c = 20$ kHz (Figures 5.10 et 5.11), cette erreur disparait logiquement : la source *trafic* étant principale, en conservant toute l'énergie sonore, l'estimation du trafic en devient meilleure. Mais dans les cas où le TIR est négatif, l'erreur augmente significativement. La distribution des erreurs relatives se focalisent alors toutes sur des niveaux sonores particuliers qui sont finalement équivalent à la somme des niveaux sonores des 2 classes de sons (équation 5.7). La distribution des erreurs relatives est alors similaire pour chaque sous-corpus et TIR avec des écarts types nuls. En prenant en compte l'ensemble des sons présents, ce filtre réalise systématiquement une surestimation des niveaux sonores *trafic*.

$$\tilde{L}_{eq,tr.,tir} = 10 \times \log_{10}\left(10^{L_{eq,tr.,tir}/10} + 10^{L_{eq,int.,tir}/10}\right) \tag{5.7}$$

En résumé, le filtre passe-bas avec une fréquence de coupure $f_c = 500$ Hz est l'approche le plus efficace sur l'ensemble du corpus sans toutefois être la plus performante sur chaque TIR. Cette méthode correspond plus à un compromis qui est fait entre l'énergie rejetée dans les TIR

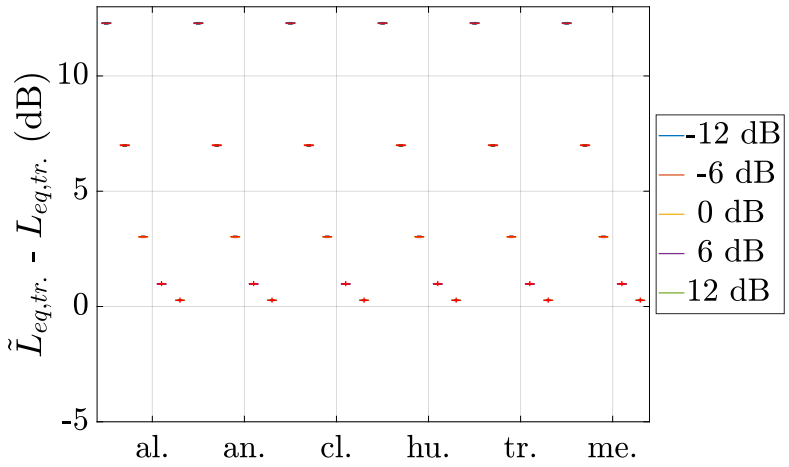


FIGURE 5.10 – Distributions des erreurs relatives entre les niveaux estimés et exactes pour chaque sous-corpus et chaque TIR pour l'estimateur filtre passe-bas à la fréquence de coupure $f_c = 20$ kHz.

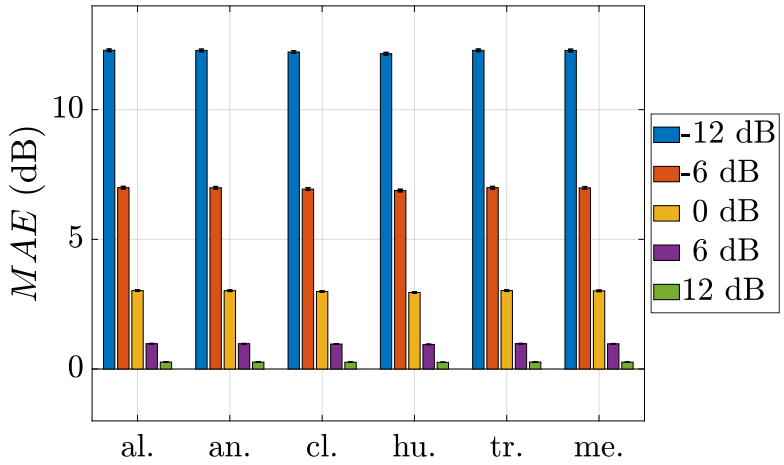


FIGURE 5.11 – Erreurs MAE pour chaque sous-corpus et chaque TIR pour l'estimateur filtre passe-bas à la fréquence de coupure $f_c = 20$ kHz.

négatifs et l'énergie conservée dans les TIR positifs.

5.5 Performances de l'estimateur basé sur la NMF

On résume les erreurs produites par les différentes versions de la NMF d'abord selon l'erreur MAE_g puis, dans un second temps, les erreurs MAE_{TIR} et MAE sont observées pour chaque méthode de NMF proposant les plus faibles erreurs.

5.5.1 Erreurs MAE_g

Le nombre de combinaisons de modalités entre les différents facteurs expérimentaux étant important, on résume, dans un premier temps, dans le Tableau 5.3, les erreurs MAE_g les plus faibles produites par la NMF SUP et SEM selon chaque valeur de β avec les modalités correspondantes. Puis, dans le cas de la NMF IS, l'influence de la représentation de la distance et du type de seuil sont observées dans les Tableaux 5.4 et 5.5.

TABLEAU 5.3 – Erreurs MAE_g les plus faibles de la NMF SUP et NMF SEM pour le corpus d'évaluation *Ambiance*, en gras-rouge, l'erreur globale la plus faible.

	f_c (kHz)	β	w_t (s)	K	MAE_g (dB)
filtre PB	20	-	-	-	4,69 (± 4,52)
	0,5	-	-	-	2,89 (± 2,84)
NMF SUP	-	0	2	50	4,29 (± 4,24)
	-	1	*all*	25	3,45 (± 3,70)
	-	2	2	25	2,84 (± 3,19)
NMF SEM	-	0	2	200	2,46 (± 1,14)
	-	**1**	**2**	**200**	**2,32 (± 1,15)**
	-	2	2	200	2,32 (± 1,26)

La NMF SUP offre des estimations moins bonnes que l'estimateur *baseline* pour $f_c = 500$ Hz à $\beta \in \{0,\ 1\}$. L'erreur la plus faible y est atteinte pour $\beta = 2$ avec une erreur similaire au filtre. Les 3 versions de la NMF SUP selon les valeurs de β obtiennent une erreur minimale pour des dictionnaires différents. Dans le cas de la divergence de KL et EUC, c'est un faible nombre d'éléments qui est requis ($K = 25$). Dans le cas de la divergence KL, le dictionnaire comprend un faible nombre d'éléments avec $w_t = all$, qui correspond au cas où le dictionnaire contient des enveloppes spectrales. La NMF SUP étant contrainte dans son fonctionnement, puisqu'elle doit avec un dictionnaire fixe *trafic* s'adapter aux différents sous-corpus, on peut supposer que ce sont ces approches qui permettent de généraliser au mieux les spectres *trafic* dans $\mathbf{W}$. Le dictionnaire de la NMF SUP avec $\beta \in \{0,\ 2\}$ est basé sur une description plus fine du corpus d'apprentissage ($w_t = 2$ s). Dans le cas de la distance EUC, étant sensible aux fortes variations d'énergie spectrale entre $\mathbf{V}_{fn}$ et $[\mathbf{WH}]_{fn}$, on suppose qu'elle privilégie un dictionnaire construit sur une description fine afin de mieux prendre en compte les variations spectrales de la source *trafic*.

La NMF SEM offre des erreurs plus faibles que l'estimateur *baseline* pour les 3 valeurs de β avec des écarts-types réduits. Dans les 3 cas, le dictionnaire est basé sur les mêmes modalités avec ici un grand nombre d'éléments ($K = 200$). La différence entre la NMF SEM et SUP réside dans la présence du dictionnaire libre $\mathbf{W_r}$ dans la NMF SEM qui permet plus d'adaptabilité (la forme des matrices $\mathbf{W_r}$ est commentée en partie 5.5.5 et est présentée en Figure 5.19).

La NMF IS étant une forme de NMF proposée pour ces travaux, plusieurs pistes sont explorées afin de trouver une configuration optimale selon le choix de la représentation de la distance

D_θ ou le type de seuillage. En considérant ces facteurs expérimentaux indépendants, il est possible de regarder leur influence séparément et ainsi éviter de calculer l'ensemble des combinaisons de modalités. Dans un premier temps, la représentation de la distance D_θ, linéaire ou exprimée à travers une fonction sigmoïde, est observée dans le Tableau 5.4 avec un seuillage dur t_h.

TABLEAU 5.4 – Erreurs MAE_g les plus faibles de la NMF IS pour le corpus d'évaluation *Ambiance* selon la représentation linéaire ou sigmoïde de la distance $D_\theta(\mathbf{W_0}\|\mathbf{W}')$.

β	w_t	K	**représentation**	t_h	MAE_g (dB)
0	500	200	linéaire	0,47	2,26 (± 2,15)
	500	200	sigmoïde	0,61	2,25 (± 2,37)
1	**500**	**200**	**linéaire**	**0,41**	**2,14 (± 2,10)**
	500	200	sigmoïde	0,60	2,14 (± 2,14)
2	500	200	linéaire	0,36	2,29 (± 2,40)
	500	200	sigmoïde	0,59	2,26 (± 2,43)

Dans un premier temps, comme pour la NMF SEM, la méthode privilégie un nombre d'éléments dans le dictionnaire important ($K = 200$). De plus, le choix de la représentation de D_θ n'influe pas sur la formation du dictionnaire puisque w_t et K restent constants. L'erreur MAE_g selon la représentation de D_θ est très peu influencée : les erreurs restent équivalentes, seuls les seuils doivent être adaptés. La fonction sigmoïde déformant D_θ, les valeurs des seuils doivent être réhaussées. L'impact de cette fonction sur l'erreur MAE_g étant très faible par rapport à la représentation linéaire, c'est cette dernière qui est donc choisie pour la suite des calculs.

L'allure moyenne de D_θ pour chaque valeur du TIR avec $w_t = 0{,}5$ s, $K = 200$ et $\beta = 1$ avec une représentation linéaire est alors résumée en Figure 5.12. Pour un seuil fixé à $t_h = 0{,}41$, le nombre d'éléments considéré dans $\mathbf{W}_{trafic}$ est donc variable. Pour les TIR négatifs, il y a plus d'éléments dont les distance D_θ sont faibles, traduisant la déviation des spectres initiaux pour des spectres liés aux classes *interférantes*. À TIR = -12 dB, c'est 106 éléments en moyenne qui sont considérés comme appartenant à la classe *trafic*. Plus le TIR augmente et plus ces distances augmentent, puisque le trafic devient la classe de son prédominante. À $TIR = 12$ dB, c'est alors 181 éléments qui sont inclus dans le dictionnaire *trafic*.

À partir de la représentation de D_θ, l'influence de la technique du seuillage sur les erreurs MAE_g est présentée dans le Tableau 5.5 pour chaque valeur de β. On rappelle que le seuillage dur définit les éléments dans le dictionnaire $\mathbf{W}'$ selon un classement binaire alors que le seuillage *firm* considère deux valeurs de seuil où les éléments dont la distance D_θ est située entre ces valeurs sont pondérés.

Là encore, la forme du dictionnaire reste la même et n'est pas influencée par le choix de la technique de seuillage. On constate que les valeurs de seuils $t_{f,1}$ et $t_{f,2}$ encadrent la valeur du seuil dur t_h. Si cet encadrement est large pour $\beta = 2$, pour $\beta \in \{0,\ 1\}$ cet encadrement est plus restreint. Cela correspond, pour $\beta = 1$, à considérer dans $\mathbf{W}_{trafic}$, en moyenne 145 (± 32) éléments issus directement de $\mathbf{W}'$ et à y inclure 8 (± 2) éléments pondérés selon la relation 3.49.

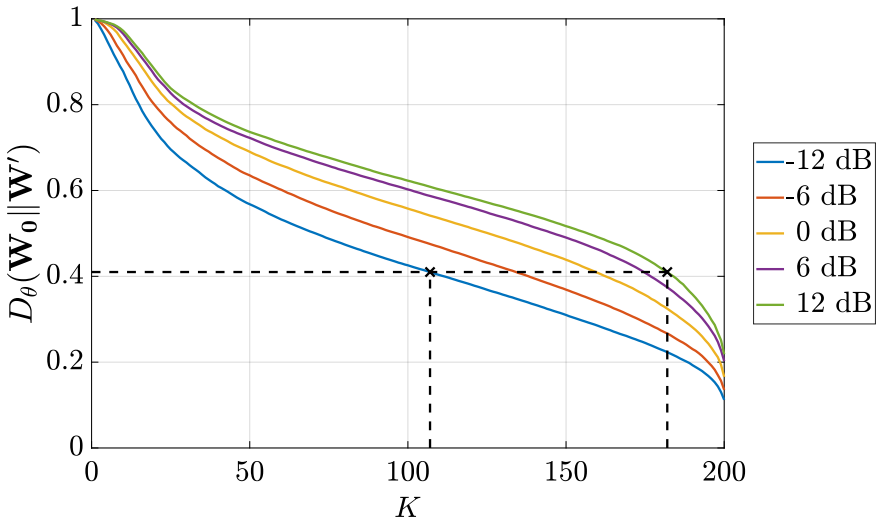


FIGURE 5.12 – Distances moyennes $D_\theta(\mathbf{W_0}\|\mathbf{W}')$ triées par ordre décroissant pour chaque TIR, obtenues pour $w_t = 0{,}5$ s, $K = 200$ et $\beta = 1$ et $t_h = 0{,}41$.

TABLEAU 5.5 – Erreurs MAE_g les plus faibles de la NMF IS pour le corpus d'évaluation *Ambiance* selon un seuillage dur ou *firm*.

β	w_t	K	t_h	$t_{f,1}$	t_{f_2}	MAE (dB)
0	500	200	0,47	-	-	2,26 (± 2,15)
	500	200	-	0,44	0,50	2,25 (± 2,14)
1	500	200	0,41	-	-	2,14 (± 2,10)
	500	**200**	-	**0,39**	**0,42**	**2,13 (± 2,16)**
2	500	200	0,36	-	-	2,29 (± 2,40)
	500	200	-	0,23	0,48	2,21 (± 2,49)

Sur l'ensemble des cas, l'apport du seuillage *firm* sur le erreurs MAE_g est très faible.

En conséquence, en vue de simplifier l'étude, on ne considère finalement que le cas d'une NMF IS avec une représentation linéaire de D_θ et un seuillage dur. De ces premiers résultats, moyennés sur l'ensemble du corpus d'évaluation, on retient donc les combinaisons de modalités les plus performantes pour chaque NMF :

— NMF SUP avec $\beta = 2$, $K = 25$, et $w_t = 2$ s,

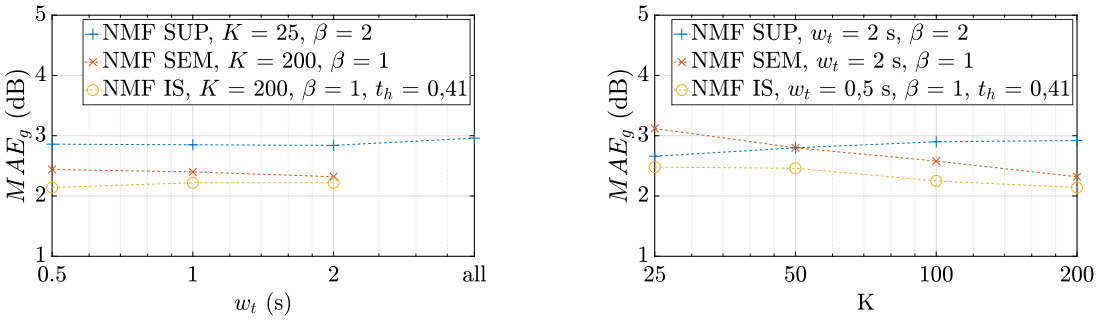
— NMF SEM avec $\beta = 1$, $K = 200$, et $w_t = 2$ s,

— NMF IS avec $\beta = 1$, $K = 200$, $w_t = 0{,}5$ s et $t_h = 0{,}41$.

Avant d'observer leur erreurs MAE_{TIR} et MAE, l'influence de la forme du dictionnaire sur les erreurs MAE_g est observée.

5.5.2 Influence des facteurs expérimentaux w_t et K

Si le choix de la méthode et de la valeur de β sont des facteurs expérimentaux prédominants dans les erreurs produites, l'influence de la forme du dictionnaire est à connaitre également. Pour cela, selon les 3 formes de NMF retenues, l'influence de la taille de la fenêtre de découpage w_t est observé en Figure 5.13(a). Pour la NMF SUP, ayant un faible nombre d'éléments dans $\mathbf{W}$ le cas où $w_t = all$ peut être observé, alors que pour la NMF SEM et IS, il n'est pas possible de l'observer puisque le nombre d'éléments ($K = 200$) est plus élevé que le nombre de fichiers audio disponibles.



(a) Influence du facteur expérimental w_t.

(b) Influence du facteur expérimental K.

FIGURE 5.13 – Influence de la forme du dictionnaire pour les 3 versions optimales de la NMF retenues.

L'influence du facteur w_t est très faible sur l'erreur de reconstruction du signal *trafic* notamment pour $w_t \in \{0.5,\ 1,\ 2\}$, pour le 3 formes de NMF. Pour la NMF IS, son influence est vouée à être réduite puisque le dictionnaire est mis à jour. On ne constate qu'un impact réel que pour le cas de la NMF SUP avec $w_t = all$ qui correspond au cas où les enveloppes des spectres *trafic* sont considérées dans le dictionnaire. L'utilisation de l'algorithme K-means et la représentation en bandes de tiers d'octave dans la construction du dictionnaire peut avoir réduit l'impact de ce facteur expérimental. La Figure 5.13(b) montre de manière synthétique l'évolution des erreurs MAE_g selon le nombre d'éléments K contenu dans le dictionnaire. Pour les 3 NMF, ce facteur est un paramètre plus influant que w_t. Les évolutions des erreurs dont d'ailleurs progressives pour chaque NMF. La NMF SUP privilégie donc bien un dictionnaire composé d'un faible nombre d'éléments en vue de mieux généraliser le dictionnaire. À l'inverse, dans le cas de la NMF IS et SEM, l'erreur diminue avec l'augmentation de K. Ces deux méthodes gagnent donc à être composées d'un grand nombre d'éléments pour mieux décrire la source sonore. Dans le cas de la NMF IS, cela se comprend d'autant plus qu'avec un nombre suffisant d'éléments dans le dictionnaire $\mathbf{W}'$, celui-ci permet une description plus fine et mieux adaptée à la scène.

5.5.3 Influence de l'initialisation de la NMF IS

Le principe de la NMF IS est d'apprendre un dictionnaire $\mathbf{W_0}$ composé de spectres de la source sonore d'intérêt pour ensuite le mettre à jour à chaque itération. La sélection des éléments

liés à la source dans $\mathbf{W'}$ se fait alors par comparaison entre le dictionnaire initial $\mathbf{W_0}$ et final $\mathbf{W'}$. Le choix d'utiliser un dictionnaire appris a pour but d'orienter les mises à jour des matrices vers la source d'intérêt. Pour visualiser l'impact de cette orientation, on compare le cas de la NMF IS optimale ($K = 200$, $w_t = 0{,}5$ s, $\beta = 1$) avec une initialisation faite par $\mathbf{W_0}$ et avec des valeurs aléatoires. Dans ce dernier cas, après 100 itérations, le dictionnaire obtenu $\mathbf{W'}$ est alors comparé à $\mathbf{W_0}$. Le Tableau 5.6 résume les erreurs MAE_g obtenues avec le seuil $t_h = 0{,}41$ et pour le seuil optimal de la NMF IS initiée avec des valeurs aléatoires.

TABLEAU 5.6 – Erreurs MAE_g les plus faibles de la NMF IS pour le corpus d'évaluation *Ambiance* selon l'initialisation du dictionnaire. En ligne 1 et 3, le seuil est celui permettant d'obtenir l'erreur la plus faible, en ligne 2, le seuil correspond à celui obtenu avec une initialisation par $\mathbf{W_0}$ mais sur une NMF IS initialisée par des valeurs aléatoires.

initialisation	β	w_t	K	t_h	MAE_g (dB)
$\mathbf{W_0}$	1	500	200	0,41	2,14 ($\pm$ 2,10)
valeurs aléatoire	1	500	200	0,41	5,21 ($\pm$ 1,31)
valeurs aléatoire	1	500	200	0,23	2,45 ($\pm$ 2,25)

L'initialisation par des valeurs aléatoires génère des erreurs supérieures à la NMF IS initiée avec $\mathbf{W_0}$. La recherche d'une erreur minimale nécessite de diminuer la valeur seuil correspondante à $t'_h = 0{,}23$. Un seuil plus faible signifie que la méthode considère un plus grand nombre d'éléments dans le dictionnaire. La Figure 5.14 résume les distances D_θ moyennes à chaque TIR pour une initialisation aléatoire. Les valeurs de D_θ sont plus faibles traduisant une disparité plus forte entre $\mathbf{W_0}$ et $\mathbf{W'}$. La dispersion des distances D_θ par TIR est également plus faible que dans le cas où $\mathbf{W_0}$ est utilisé (voir Figure 5.12). On constate également qu'aucune valeur de D_θ n'atteint une similarité de 1. Sans l'initialisation, le dictionnaire mis à jour s'éloigne donc plus de la source *trafic* générant des erreurs plus élevées. Utiliser $\mathbf{W_0}$ à la première itération a donc un impact réel et permet de mieux focaliser les mises à jour du dictionnaire sur la source *trafic* et ainsi de mieux reconstruire cette source sonore.

5.5.4 Erreurs MAE_{TIR} et fonctions de coût

La NMF est réalisée en cherchant à minimiser la fonction de coût $D(\mathbf{V}\|\mathbf{WH})$ et se base donc sur la qualité de la reconstruction du signal global dans lequel la composante *trafic* est présente. En conséquence, la NMF réduit la distance entre $\mathbf{V}$ et $\mathbf{WH}$ sans pour autant assurer que la restitution des différentes composantes présentes dans le signal audio soit bonne. Il est donc utile de mettre en parallèle l'évolution de cette fonction de coût avec celle des erreurs MAE_g. Pour cela, la Figure 5.15 résume l'évolution de la fonction de coût $D(\mathbf{V}\|\mathbf{WH})$ moyennée sur l'intégralité du corpus et des erreurs MAE_g pour chacune de ces méthodes.

La convergence de la NMF IS est meilleure que les deux autres NMF, ce qui est cohérent puisque la matrice $\mathbf{W}$ est mise à jour, ce qui facilite la décroissance de $D(\mathbf{V}\|\mathbf{WH})$ et ainsi améliore la rapidité de convergence. La NMF SEM est, quant à elle, celle dont la fonction de coût est la plus importante. Pour la NMF SUP, les valeurs de la fonction de coût sont faibles et

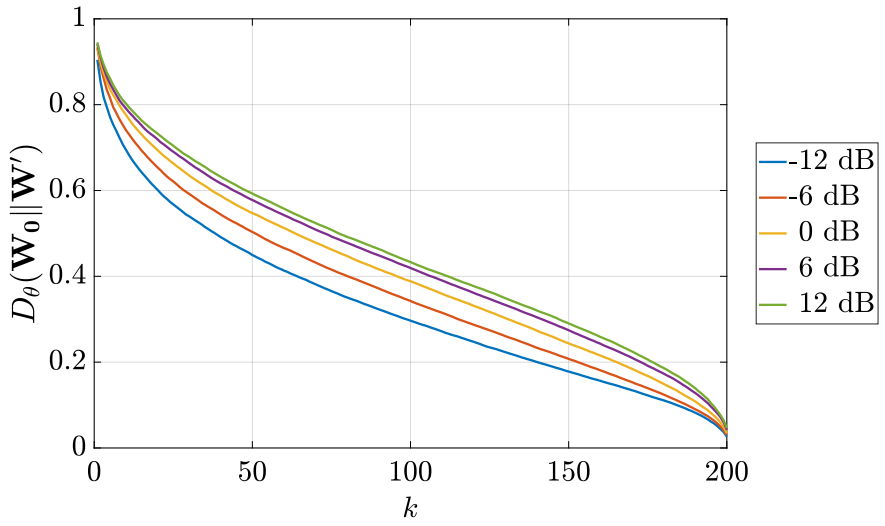
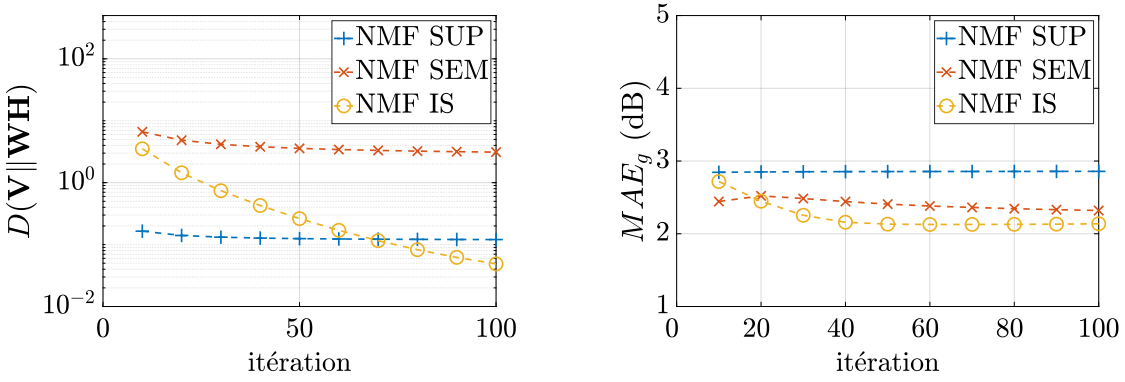


FIGURE 5.14 – Distances moyennes $D_\theta(\mathbf{W_0}\|\mathbf{W}')$ triées par ordre décroissant pour chaque TIR, obtenues pour $w_t = 0{,}5$ s, $K = 200$ et $\beta = 1$ pour un dictionnaire initialisé par des valeurs aléatoires.



(a) Fonctions de coût moyennées.

(b) Erreurs MAE_g moyennées.

FIGURE 5.15 – Evolution de la fonctions de coût $D(\mathbf{V}\|\mathbf{WH})$ et de l'erreur MAE_g moyennes pour les combinaisons optimales des NMF SUP, SEM et IS sur l'intégralité du corpus d'évaluation *Ambiance*.

convergent rapidement. N'ayant qu'une matrice à mettre à jour à partir du dictionnaire fixe, la matrice $\mathbf{H}$ trouve rapidement une forme optimale qui minimise la fonction de coût. À l'inverse pour la NMF SEM et IS où plusieurs matrices sont à mettre à jour, respectivement $\mathbf{W_r}$, $\mathbf{H_r}$, $\mathbf{H_r}$ et $\mathbf{W}'$ et $\mathbf{H}$, ce qui nécessité plus d'itérations. On constate que la fonction de coût pour la NMF IS, bien que plus faible, n'a pas convergé vers une valeur fixe au bout de 100 itérations. Il semble donc possible d'améliorer la similarité entre $\mathbf{V}$ et $\mathbf{WH}$. Toutefois, ce comportement est à mettre en parallèle avec l'évolution de l'erreur MAE_g en Figure 5.15(b). On constate qu'au bout de 100 itérations, la valeur de l'erreur moyenne pour la NMF IS est quasi constante : entre l'erreur à l'itération 90 et 100, on améliore l'erreur de 0,2 %. En conséquence, même si la fonction de coût

n'a pas convergé à une valeur fixe, l'erreur relative au signal *trafic* devient quasiment constante après 100 itérations. Les mises à jour de $\mathbf{W}'$ et $\mathbf{H}$ modélisent alors les sources interférantes et non plus le trafic.

De ces 3 estimateurs, leurs erreurs MAE_{TIR} sont exprimées dans le Tableau 5.7. À ces résultats sont ajoutés ceux des estimateurs *baseline* pour $f_c = 500$ Hz et $f_c = 20$ kHz. En gras-rouge, les erreurs MAE_{TIR} les plus faibles, en gras-noir les erreurs obtenues par les NMF qui sont inférieures à l'erreur de la baseline $f_c = 500$ Hz.

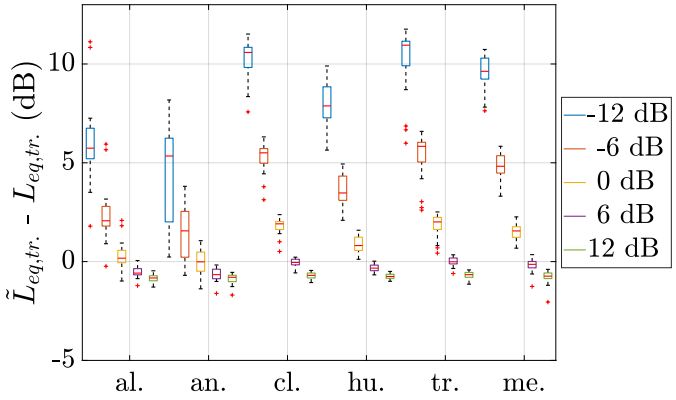
TABLEAU 5.7 – Erreurs MAE_{TIR} en dB selon les combinaisons optimales de la NMF SUP, SEM et IS avec l'estimateur *baseline* $f_c = 500$ Hz et les erreurs pour le filtre passe-bas $f_c = 20$ kHz.

méthode	**-12**	**-6**	**0**	**6**	**12**
filtre PB, $f_c = 20$ kHz	12,25 (± 0,05)	6,96 (± 0,05)	3,00 (± 0,03)	0,97 (± 0,01)	**0,26 (± 0,00)**
filtre PB, $f_c = 0{,}5$ kHz	7,39 (± 3,00)	3,44 (± 1,65)	**1,17 (± 0,24)**	1,03 (± 0,26)	1,45 (± 0,13)
NMF SUP	8,08 (± 2,44)	3,84 (± 1,58)	1,15 (± 0,62)	**0,35 (± 0,20)**	**0,77 (± 0,07)**
NMF SEM	**2,98 (± 2,11)**	**1,52 (± 0,60)**	1,60 (± 0,47)	2,49 (± 0,30)	3,02 (± 0,22)
NMF IS	**5,22 (± 2,62)**	**2,72 (± 1,24)**	1,26 (± 0,35)	**0,75 (± 0,34)**	**0,83 (± 0,23)**

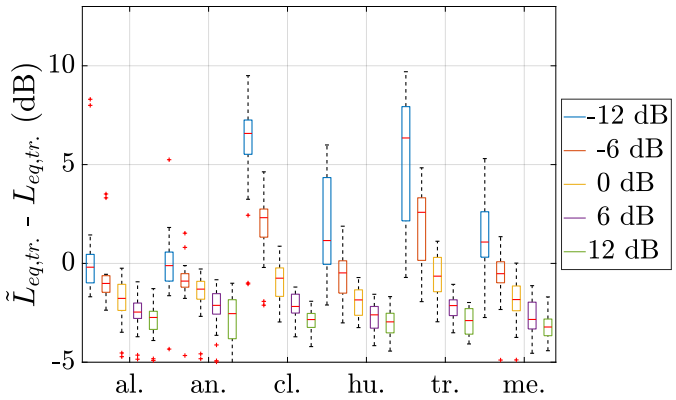
De la même manière que pour l'estimateur filtre dans le Tableau 5.2, la méthode la plus efficace en moyenne sur l'ensemble du corpus n'est pas forcément la méthode la plus efficace pour chaque TIR. Ici, même si elle est pour quatre valeurs du TIR inférieure à la baseline, la NMF IS n'est jamais la méthode qui propose l'erreur la plus faible. Pour $TIR \in \{-12, -6\}$ dB, la NMF SEM est la plus performante, alors que pour $TIR \in \{6, 12\}$ dB, la NMF SUP supplante les deux autres méthodes. Le cas où $TIR = 0$ dB est un cas unique où aucune NMF n'améliore les résultats de la baseline. Pour $TIR = 12$ dB c'est finalement le filtre $f_c = 20$ kHz (c'est-à-dire l'absence de filtrage), qui se trouve être la méthode la plus performante. Lorsque, le niveau du trafic est très important, sa meilleure estimation est le niveau global de la mixture. Il vaut donc mieux ne rien faire sur le signal de mixture que d'appliquer la NMF. Ce résultats peut paraître décevant mais le problème est que, sur un problème concret, on ne connait pas le niveau à priori du trafic. Toutefois, la NMF IS est la seule méthode à être systématiquement inférieure à la baseline (hormis à $TIR = 0$ dB) et même si elle n'est pas systématiquement la plus performante, elle est celle qui s'adapte le mieux aux différentes valeurs du TIR. Les raisons du comportement des 3 NMF seront étudiés dans la partie suivante.

5.5.5 Erreurs MAE pour chaque ambiance et valeur du TIR

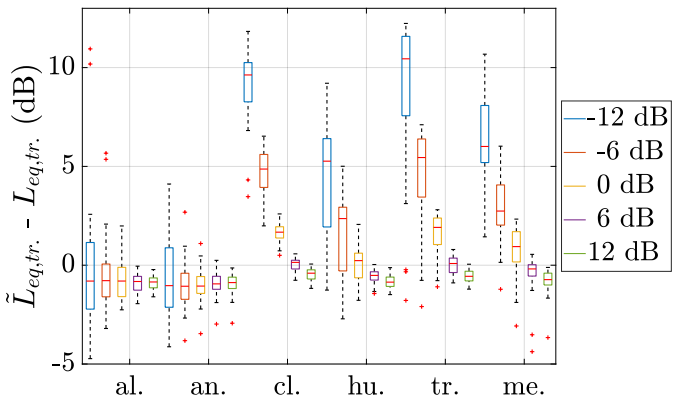
Après avoir observé l'erreur globale MAE_g, pour déterminer les combinaisons de modalités qui proposent les erreurs moyennes les plus faibles, et l'erreur moyenne selon chaque TIR, la distribution des erreurs relatives et l'erreur MAE sont maintenant observées. Ces erreurs sont calculées à partir des différences entre les niveaux sonores exacts, $L_{eq,tr.}$, et estimés, $\tilde{L}_{eq,tr.}$ sur



(a) Erreurs relatives pour la NMF SUP.

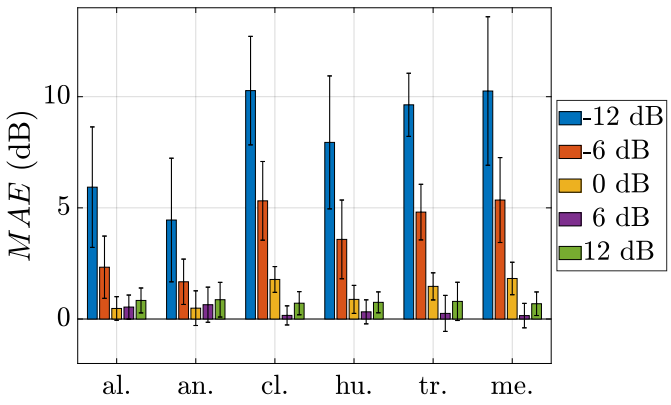


(b) Erreurs relatives pour la NMF SEM.

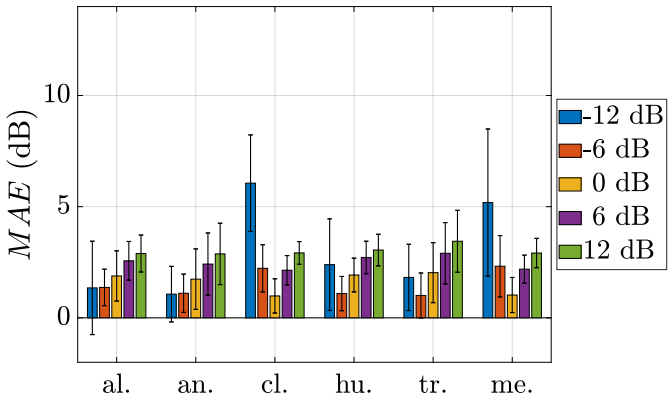


(c) Erreurs relatives pour la NMF IS.

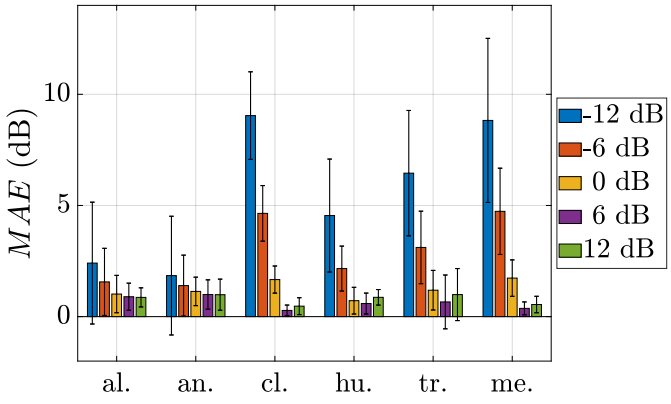
FIGURE 5.16 – Distributions des erreurs relatives pour chaque classe pour la NMF SUP (a), SEM (b) et IS (c).



(a) Erreurs MAE pour la NMF SUP.



(b) Erreurs MAE pour la NMF SEM.



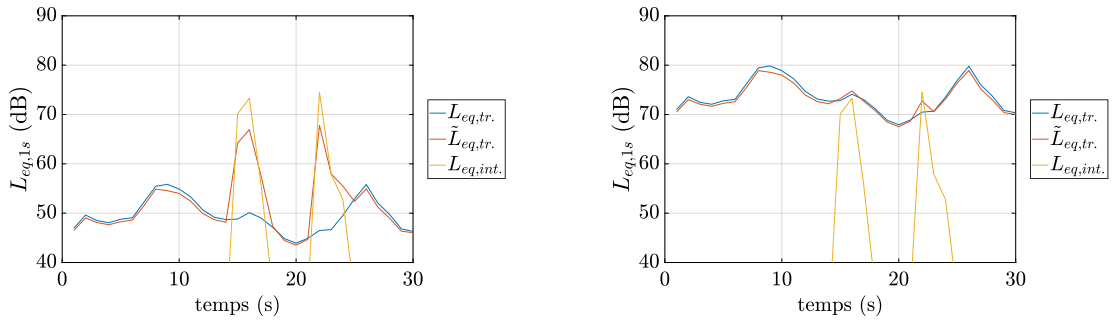
(c) Erreurs MAE pour la NMF IS.

FIGURE 5.17 – Erreurs MAE pour chaque classe pour la NMF SUP (a), SEM (b) et IS (c).

l'ensemble des 25 scènes pour chaque valeur de TIR et sous-corpus.

Dans les Figures 5.16 et 5.17, la distribution des erreurs relatives et les erreurs MAE des 3 méthodes NMF optimales pour chaque TIR et chaque ambiance sont exposées. Les Figures 5.16(a) et 5.17(a) permettent de visualiser les performances très variables de la NMF SUP. Dans le cas où $TIR \in \{-12, -6\}$ dB, et cela pour les 6 sous-corpus, la NMF SUP obtient de fortes erreurs MAE. Les plus fortes erreurs sont obtenues pour les sous corpus *climat*, *humains*, *transport* et *mécanique* car on y trouve les classes de sons interférantes dont les allures spectrales sont les plus similaires à celles du trafic. Même pour les sous-corpus *alerte* et *animaux*, la méthode échoue à obtenir de faibles erreurs. Ces performances sont toutefois contre-balancées par des erreurs beaucoup plus faibles dans les TIR positifs, notamment à $TIR = 6$ dB. La présence exclusive d'élément *trafic* dans le dictionnaire sur des scènes où cette classe de son est prépondérante rend la NMF plus performante quel que soit le sous-corpus. De la même manière que le filtre à 500 Hz, la NMF SUP surestime les niveaux sonores *trafic* pour les TIR négatifs et les sous-estime lorsque celui-ci est positif.

On illustre le comportement de la NMF SUP dans le cas d'une scène de l'ambiance *alerte* pour TIR = -12 dB et TIR = 12 dB dans la Figure 5.18 où les évolutions du $L_{eq,1s}$ des signaux *trafic* exact, *trafic* estimé par la NMF SUP optimale et du signal *interférant* sont tracées. Dans le cas où le TIR est négatif, lorsque le signal *interférant* est émergent, on observe que le signal *trafic* estimé inclut ce signal. N'étant composé que d'éléments *trafic* dans le dictionnaire c'est donc que les bases *trafic* de $\mathbf{W}$ sont activées pour modéliser un signal qui ne l'est pourtant pas. Ce comportement disparait toutefois pour $TIR = 12$ dB où le signal *trafic* est correctement modélisé sans être impacté par la classe de son interférante.

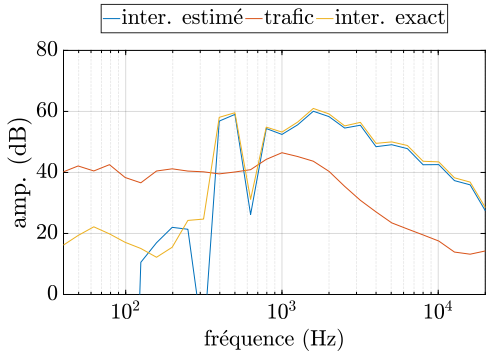


(a) Comparaison du niveau sonore pour TIR = -12 dB pour la scène 25 du sous-corpus *alerte*.

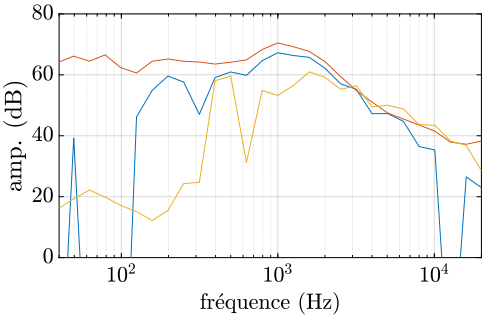
(b) Comparaison du niveau sonore pour $TIR = 12$ dB pour la scène 25 du sous-corpus *alerte*.

FIGURE 5.18 – Comparaisons des niveaux sonores équivalents des classes *trafic* et *interférante* pour 2 valeurs du TIR (-12 dB et 12 dB) pour la scène 25 du sous-corpus *alerte*.

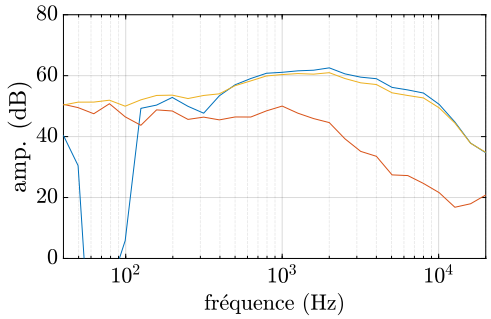
La NMF SEM (Figures 5.16(b) et 5.17(b)) présente un comportement différent de la NMF SUP : les erreurs relatives sont notamment négatives et les erreurs MAE sont plus faibles lorsque les valeurs du TIR sont négatives, hormis pour les sous-corpus *climat* et *mécanique*. L'ajout de $\mathbf{W_r}$ dans le dictionnaire est donc déterminant puisque c'est cet élément qui différencie les deux méthodes. Cependant, lorsque le TIR devient positif et le trafic dominant, l'erreur augmente sys-



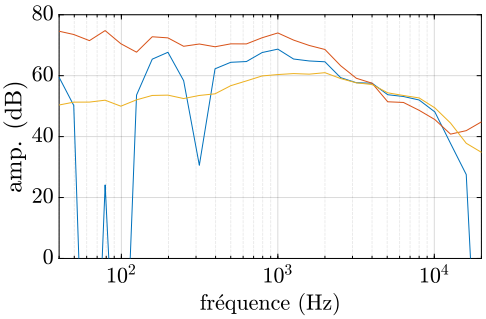
(a) Comparaison pour TIR = -12 dB pour la scène 2 du sous-corpus *alerte*.



(b) Comparaison pour TIR = 12 dB pour la scène 2 du sous-corpus *alerte*.



(c) Comparaison pour TIR = -12 dB pour la scène 3 du sous-corpus *climat*.



(d) Comparaison pour TIR = 12 dB pour la scène 3 du sous-corpus *climat*.

FIGURE 5.19 – Comparaisons des spectres des classes *trafic* et *interférante* avec la somme des 2 éléments de $\mathbf{W_r}$ pour 2 valeurs du TIR (-12 dB et 12 dB) pour 2 sous-corpus (*alerte* et *climat)*.

tématiquement là où les erreurs relatives décroissent traduisant une sous-estimation des niveaux sonores par la NMF SEM. On compare dans les Figures 5.19, le spectre du signal interférant recomposé par le produit $\mathbf{W_r H_r}$ obtenue pour une scène *alerte* et *climat* avec les spectres des signaux *trafic* et *interférant* à $TIR \in \{-12, 12\}$ dB. Dans les cas où TIR = -12 dB, le spectre du signal interférant estimé correspond bien à celui de la classe *interférante*. Sa modélisation par la NMF SEM dans les éléments libres est donc correcte laissant la partie *trafic* au dictionnaire $\mathbf{W_s}$. À l'opposée, pour TIR = 12 dB, le spectre estimé diverge de celui de la classe *interférante*. Certaines parties des spectres participent alors à la modélisation de la composante *trafic*, ce qui réduit la justesse de son estimation du niveau sonore. Les degrés de liberté de la NMF SEM sont donc un avantage lorsque le trafic est peu présent car ils permettent bien d'intégrer la classe de son *interférante*. Mais cette liberté joue en sa défaveur lorsque le trafic devient la classe de son prédominante et où ses composantes sont incluses dans $\mathbf{W_r}$.

Enfin, les Figures 5.16(c) et 5.17(c) sont dédiées à la NMF IS optimale. Si l'évolution des erreurs MAE de la NMF IS présentent des allures similaires à celles de la NMF SUP (erreurs plus fortes pour les TIR négatifs que dans les TIR positifs) mais avec des erreurs moindres

notamment pour les TIR négatifs, la distribution des erreurs relatives est différente notamment pour les sous-corpus *alerte* et *animaux*. Pour ces deux sous-corpus, l'ensemble des erreurs sur les 5 valeurs du TIR sont négatives. La taille des boites est également plus larges sur les autres sous-corpus traduisant une variance plus grande des erreurs. Toutefois les erreurs pour les sous-corpus *climat*, *transport* et *mécanique* restent élevées en raison de la proximité de leur spectre avec ceux du trafic. À partir de $TIR \geq 0$ dB, les erreurs deviennent faibles ($MAE < 1,7$ dB).

La valeur seuil $t_h = 0{,}41$ est donc la valeur optimale permettant une erreur minimale sur l'intégralité du corpus, mais selon le sous-corpus ou la valeur du TIR ce seuil est susceptible de varier. Dans les Figures 5.20, l'évolution de l'erreur MAE est ainsi tracée en fonction du seuil t_h pour chaque sous-corpus et pour 3 valeurs de TIR.

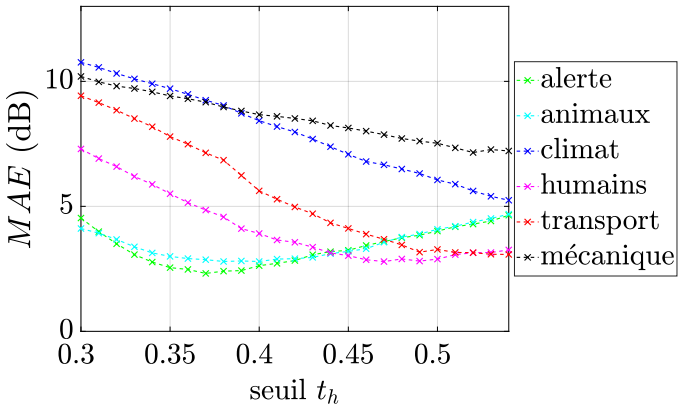
— Pour TIR = -12 dB, l'erreur MAE minimale diffère selon les sous-corpus. Pour les classes de son *transport* et *humains*, leur seuil optimal correspondant à l'erreur minimale est situé vers 0,50. Dans le cas du *climat* et de *mécanique*, ce seuil semble se situer au delà de la plage de variation testée. Pour ces 4 sources, dont les erreurs à ces TIR sont les plus fortes, leur spectre étant similaire à celui du *trafic*, il y a intérêt à augmenter la valeur du seuil t_h pour restreindre le nombre d'éléments dans $\mathbf{W}_{trafic}$. À l'inverse pour les classes *animaux* et *alerte*, dont l'allure spectrale est différente de celle du trafic, le seuil optimal peut être diminué. Ainsi, la valeur seuil optimale pour TIR = -12 dB peut être relevée à $t_h = 0,49$, sur l'ensemble des sous-corpus pour une erreur $MAE_{-12} = 4{,}61$ ($\pm$ 1,91) dB avec alors un nombre moyen d'éléments $K = 79$ ($\pm$ 22).

— Lorsque $TIR = 0$ dB, le trafic devenant prédominant, le dictionnaire $\mathbf{W}'$ mis à jour contient plus d'éléments trafic. En diminuant le seuil t_h, plus d'éléments relatif à cette source sonore sont intégrés dans $\mathbf{W}'$. Les classes de sons *climat* et de *mécanique* restent encore en marge avec un seuil optimal élevé situé entre 0,45 et 0,50.

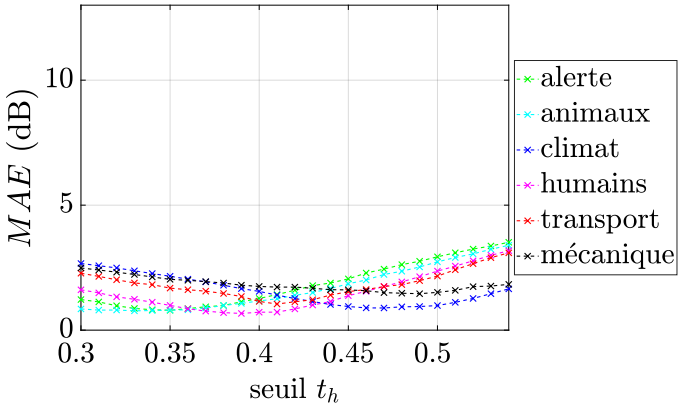
— Enfin, pour $TIR = 12$ dB, pour un seuil optimal $t_h = 0{,}31$, un plus grand nombre d'éléments du dictionnaire est considéré ($K_{moyen} = 142$ ($\pm$ 22)), l'erreur MAE_{12} diminue alors à 0,20 ($\pm$ 0,08) dB, soit même une erreur inférieure au filtre passe-bas à 20 kHz (correspondant à la mixture sonore complète).

5.6 Conclusion du chapitre

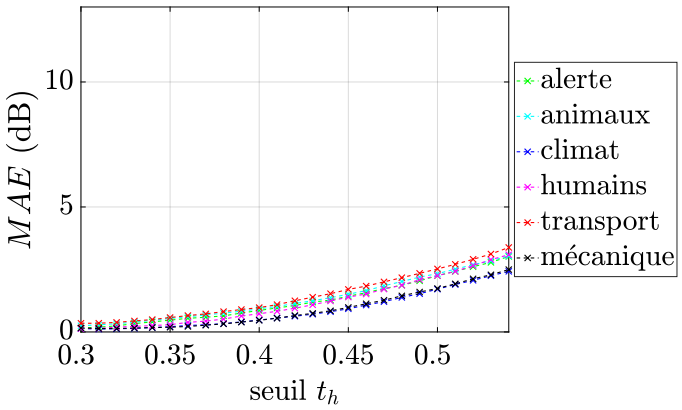
Cette première étude, à partir du corpus élémentaire *Ambiance*, a permis d'établir le fonctionnement de plusieurs formes de NMF face à des scènes sonores urbaines. Ces résultats révèlent la difficulté à obtenir une méthode efficace et performante quelles que soient les différentes sources sonores interférantes ou la présence variable du trafic. La NMF SUP, composée d'un dictionnaire comprenant des éléments relatifs au trafic, se révèle être peu performante lorsque le trafic est peu présent en activant des bases *trafic* pour simuler d'autres sources sonores mais est plus efficace lorsque les valeurs du TIR sont positives. La NMF SEM, à l'inverse, est très efficace pour des



(a)



(b)



(c)

FIGURE 5.20 – Évolution de l'erreur MAE pour chaque sous-corpus selon la valeur seuil t_h pour $TIR = -12$ dB (5.20(a)), $TIR = 0$ dB (5.20(b)) et $TIR = 12$ dB (5.20(c)).

TIR négatifs, mais échoue à définir correctement le signal *trafic* lorsque cette source sonore devient prédominante. L'ajout du dictionnaire libre $\mathbf{W_r}$ est alors un atout ou un inconvénient selon

la prédominance du trafic : dans le premier cas $\mathbf{W_r}$ modélise convenablement la classe de son *interférante* ce qui permet une bonne estimation du niveau sonore du trafic, mais dans le second cas, cet élément du dictionnaire inclut des composantes *trafic* ce qui détériore son estimation. Finalement, la NMF IS se trouve être l'approche qui minimise le mieux l'erreur sur l'ensemble du corpus. Elle réalise un compromis entre les deux méthodes : elle offre suffisamment de liberté par la mise à jour de $\mathbf{W_0}$ pour s'adapter aux différentes ambiances et sources sonores tout en étant contrainte par le seuillage à ne conserver que les éléments les moins divergents. Les différentes pistes étudiées ont permis d'établir que la simple estimation de la distance $D_\theta(\mathbf{V}\|\mathbf{WH})$ avec l'extraction de la composante *trafic* par seuillage dur est l'approche la plus efficace. À la différence des deux autres méthodes, la NMF IS présente l'avantage de modéliser la composante trafic telle qu'elle est capté par le capteur et donc susceptible de prendre en compte l'impact de l'environnement. Si on se réfère au problème posé au chapitre 1 dans la partie 1.1, la NMF IS revient à estimer directement le signal $S(t)$ où la composante *trafic*, $S_{tr.}(t)$, est ensuite extraite par la méthode de seuillage dur. La NMF SUP et SEM, quant à elles, tente de déterminer $S_{tr.}(t)$ à partir des connaissances apprises sur la source trafic et donc $s_{tr.}(t)$. La NMF IS présente donc l'intérêt de mieux considérer l'impact de l'environnement sur la source sonore que les deux autres méthodes et d'être ainsi plus généralisable.

Chapitre 6

Performances de la NMF sur le corpus d'évaluation *SOUR*

Résumé

Ce dernier chapitre résume les performances de la NMF appliquée sur le corpus d'évaluation *SOUR*, assimilable à des enregistrements sonores urbains. La meilleure combinaison obtenue est celle constituée par la NMF IS pour la distance Euclidienne avec 300 éléments dans le dictionnaire, une trame temporelle $w_t = 1$ seconde et un seuil dur $t_h = 0{,}35$. Cette version peut alors être considérée sur des mesures faites en ville afin d'estimer le niveau sonore du trafic routier. Plusieurs pistes sont enfin explorées afin d'améliorer ces performances faisant intervenir des contraintes de régularité temporelle ainsi que l'adaptation du seuil de la NMF IS en fonction de l'environnement géographique.

Au chapitre précédent, les aptitudes de la NMF à déterminer le niveau sonore du trafic routier au sein de mixtures sonores urbaines ont été étudiées avec l'aide du corpus élémentaire *Ambiance*. Si cette partie a permis de démontrer l'intérêt de la NMF IS sur de tels environnements sonores, les performances et les combinaisons obtenues ne correspondent toutefois pas à celles qui proviendraient d'enregistrements sonores urbains. En effet, ce corpus présente une artificialité (classe de son interférante spécifique, calibration des niveaux sonores du trafic) qui n'est pas représentative de l'ESU. Ainsi, dans ce chapitre, la NMF est appliquée sur un second corpus, le corpus d'évaluation de Scènes sOnores Urbaines Réalistes (abrégé *SOUR*). L'aspect « réaliste » de ce corpus provient de la construction des scènes sonores qui a été basée sur des enregistrements réels effectués en ville. L'approche qui obtient les erreurs les plus faibles pourra alors être retenue pour des applications sur des cas réels. Dans un premier temps, un rappel du corpus, des facteurs expérimentaux, de leurs modalités respectives et du calcul de métrique est réalisé. Puis les erreurs générées par l'estimateur *baseline* et par la NMF sont présentées. Des pistes visant à améliorer

ces estimations sont ensuite étudiées et proposées.

6.1 Rappel de l'expérience menée

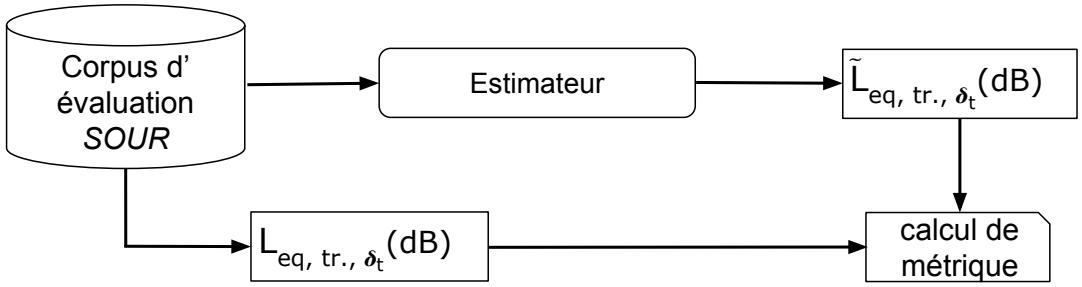


FIGURE 6.1 – Schéma-bloc des étapes dans l'estimation du niveau sonore du trafic pour le corpus d'évaluation *SOUR*.

De prime abord, les étapes mises en place dans cette partie sont similaires à celles du chapitre précédent et sont présentées dans la Figure 6.1 : les scènes sonores du corpus d'évaluation *SOUR* sont soumises à un estimateur qui détermine le niveau sonore du trafic estimé, $\tilde{L}_{eq,tr.,\delta_t}$ qui est comparé à sa valeur exacte, $L_{eq,tr.,\delta_t}$ par le calcul d'une métrique. Le corpus d'évaluation *SOUR*, présenté en détail dans la partie 4.5, est composé de 74 fichiers audio d'une durée totale de 2h50, divisés en 4 ambiances sonores : *Parc* (8 scènes, abrégé *Pa.*), *Rue calme* (35 scènes, abrégé *Ca.*), *Rue bruyante* (23 scènes, abrégé *Br.*), *Rue très bruyante* (8 scènes, abrégé *TrBr.*). Ce corpus a été réalisé à partir d'enregistrements sonores et de leurs annotations afin d'obtenir des scènes sonores simulées dont la structure temporelle est similaire à celle des scènes réelles. La qualité et le réalisme de ces scènes ont été validés grâce à un test perceptif. Ces scènes étant assimilables à des enregistrements sonores, elles permettent d'évaluer dans des conditions réalistes les performances de la NMF face à de telles mixtures sonores. Comparées au corpus *Ambiance*, ces scènes présentent une part du trafic plus importante et mélangent de multiples sources sonores (aboiement de chien, sifflement d'oiseaux, voix, bruit de rue, bruit de pas, sonnerie, portes...). Les sources sonores appartenant aux classes de sons interférantes *climat* (vent, pluie, orage) et *transport* (tramway, avion et train) ne sont pas présentes dans ce corpus.

Chaque scène du corpus est soumise à un estimateur qui détermine le niveau sonore du trafic routier. Le premier estimateur *baseline* reste le même que dans le chapitre 5 : un filtre passe-bas (abrégé filtre PB) de fréquence de coupure f_c avec $f_c \in \{100,\ 500,\ 1k,\ 2k,\ 5k,\ 10k,\ 20k\}$ Hz. Son application reste similaire au cas du chapitre précédent : le spectrogramme du signal audio est coupé à la fréquence f_c et toute l'énergie située dans la bande passante est assimilée au trafic routier (voir Figure 5.2). Le second estimateur est basé sur plusieurs formes de NMF : NMF supervisée (NMF SUP), semi-supervisée (NMF SEM) ainsi que la NMF initialisée seuillée (NMF IS). Le choix de la β-divergence reste circonscrit à $\beta \in \{0,\ 1,\ 2\}$. L'apprentissage du diction-

TABLEAU 6.1 – Facteurs expérimentaux et leurs modalités utilisés pour le corpus *SOUR*.

Facteur expérimentaux	**Modalités**				**Nombre de modalités**
Environnement sonore	parc 'Pa.'	rue calme 'Ca.'	rue bruyante 'Br.'	rue très bruyante 'TrBr.'	4
Méthode	filtre PB	NMF SUP	NMF SEM	NMF IS	4
f_c (kHz)	1 0.5	1 2	5 10	20	7
w_t (s)	0.5	1	2	*all*	4
K	25	50 100	200	300	5
β	0	1	2		3
seuillage dur t_h	de 0.20 à 0.60 avec un pas de 0.01				41

naire suit les même étapes qu'au chapitre 5 : chaque spectrogramme, issu des 53 fichiers audio constituant le corpus d'apprentissage, est découpé en trames temporelles de durées variables $w_t \in \{0.5,\ 1,\ 2\}$ seconde(s). Les valeurs *rms* sont ensuite calculées selon la fréquence. L'algorithme de clustering K-means est ensuite appliqué afin de fixer le nombre d'éléments dans $\mathbf{W}$ à $K \in \{25,\ 50,\ 100,\ 200,\ 300\}$. Le cas *all*, où la valeur *rms* est calculée sur l'intégralité du corpus, est aussi conservé avec la réduction des matrices toujours restreinte à $K_{w_t=all} = \{25,\ 50\}$. Les dictionnaires, constitués d'éléments *trafic*, composent les dictionnaires $\mathbf{W}$ de la NMF SUP, les dictionnaires $\mathbf{W_s}$ de la NMF SEM et les dictionnaires initiaux $\mathbf{W_0}$ de la NMF IS qui sont ensuite mis à jour. Pour la NMF SEM, le nombre d'éléments dans le dictionnaire $\mathbf{W_r}$ est maintenu à $J = 2$. Dans le cas de la NMF IS, l'influence de l'opérateur sigmoïde dans le calcul de la distance $D_\theta(\mathbf{W_0}\|\mathbf{W}')$ et l'utilisation du seuil *firm* s'étant relevées très faibles sur le corpus d'évaluation *Ambiance*, l'étude se réduit au seul cas de la représentation linéaire de la distante D_θ avec un seuillage dur de seuil t_h. Comme les scènes sonores sont plus longues (de 1 à 4 minutes), le nombre d'itération est étendu à 200 pour toutes les combinaisons testées. Le résumé de ces facteurs expérimentaux et de leurs modalités respectives se trouve dans le Tableau 6.1. Dans le cas de l'estimateur *baseline*, 28 associations de modalités sont réalisées (4 environnements sonores $\times$ 7 f_c). Pour l'estimateur NMF, c'est en tout 13 776 combinaisons qui sont possibles (4 environnements sonores $\times$ 3 β $\times$ (3 w_t $\times$ 4 K + 1 w_t $\times$ 2 $K_{w_t=all}$) $\times$ (2 méthode+ 41 t_h)).

En sortie de l'estimateur est calculé un niveau sonore équivalent du trafic, $\tilde{L}_{eq,tr.,\delta_t}$. Dans le précédent corpus, comme chaque scène possédait une durée similaire de 30 secondes, les niveaux sonores équivalent étaient calculés pour une durée d'intégration δ_t de 30 secondes aussi. Ici, la durée des scènes dans le corpus *SOUR* est variable, il serait donc moins rigoureux de comparer entre eux les niveaux sonores équivalents des différentes scènes intégrés sur leurs durées totales si celles-ci sont différentes. La scène la plus courte dure 55 secondes (scène n° 14 de l'ambiance *Rue bruyante*) et la plus longue dure 270 secondes (scène n° 4 de l'ambiance *Rue bruyante* également). Pour harmoniser le calcul de la métrique, c'est un niveau sonore équivalent qui est calculé avec une durée d'intégration δ_t de 60 secondes :

$$L_{eq,tr.,60s} = 10 \times \log_{10}\left(\frac{1}{60}\int_{t_{init}}^{t_{init}+60} 10^{L_{p,tr.}(t)/10} dt\right) \tag{6.1}$$

avec $L_{p,tr.}(t)$ l'expression du niveau sonore. Pour cela, dans chaque ambiance sonore, les signaux *trafic* obtenus en sortie de l'estimateur sont concaténés les uns aux autres. Puis le niveau sonore équivalent est successivement calculé toutes les 60 secondes. Il est alors possible de déterminer un niveau sonore $L_{eq,tr.,60s}$ dont une partie est calculée sur une scène et l'autre partie sur la scène suivante. Dans le cas des dernières secondes du dernier signal, si la durée restante est inférieure à 30 secondes, le signal est intégré au précédent. Si sa durée est supérieure à 30 secondes, le signal est considéré comme un signal à part entière. On résume, en Figure 6.2, un exemple du procédé. La calibration réalisée dans la partie 4.5.4 en vue d'harmoniser les signaux permet ici d'éviter des changements trop brusques entre les scènes et ainsi d'être moins sensible aux variations de l'énergie sonore d'une scène à l'autre.

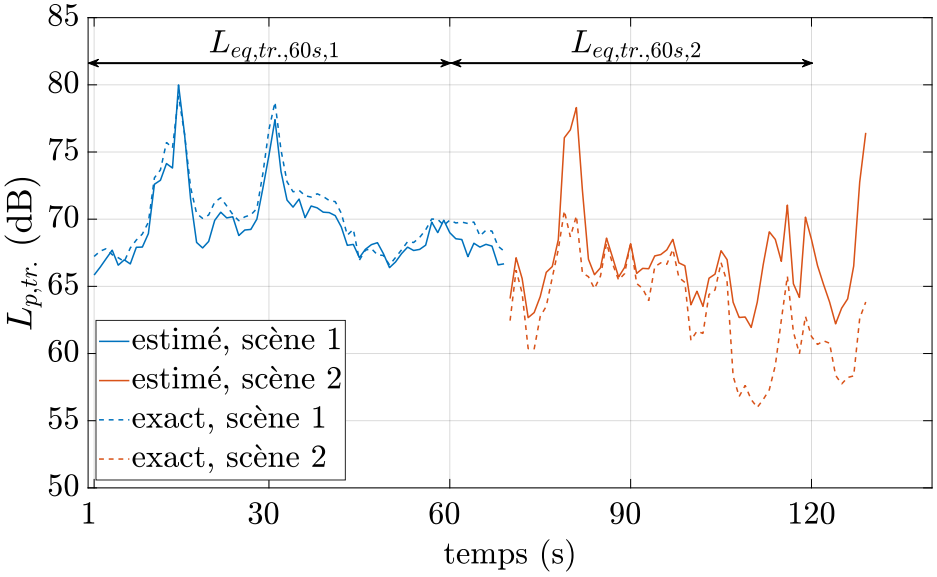


FIGURE 6.2 – Estimation des niveaux sonores équivalents pour une durée d'intégration de 60 secondes sur deux scènes. Le deuxième niveau sonore $L_{eq,tr.,60s,2}$ inclut 9 secondes de la scène 1 et 51 secondes de la scène 2.

Le nombre d'erreurs calculées par ambiance est donc différent du nombre de scènes qui la compose. Le Tableau 6.2 récapitule le nombre de scènes par ambiance ainsi que leur durée cumulée et le nombre de niveaux sonores équivalent à la minute calculé en conséquence (N_{60}).

L'erreur MAE_{60} est calculée pour chaque association de modalité pour chaque ambiance sonore :

$$MAE_{60} = \frac{\sum_{i=1}^{N_{60}} |L_{eq,tr.,60s,i} - \tilde{L}_{eq,tr.,60s,i}|}{N_{60}} \tag{6.2}$$

Cette erreur est ensuite moyennée sur l'intégralité des ambiances sonores pour déterminer la combinaison de modalités optimale qui permet d'obtenir l'erreur MAE_g la plus faible sur

TABLEAU 6.2 – Corpus d'évaluation *SOUR* par ambiance sonore selon le nombre de scènes N, leur durée totale et le nombre de niveaux sonores calculées N_{60}.

Ambiance sonore	**Parc**	**Calme**	**Bruyant**	**Très Bruyant**	**Total**
nombre de scènes **N**	8	35	23	8	74
durées (s)	960	4636	3366	1285	10247
$\mathbf{N_{60}}$	16	78	57	22	173

l'ensemble du corpus :

$$MAE_g = \frac{\sum_{j=1}^{4} MAE_{60,j}}{4}. \tag{6.3}$$

Le choix a été fait de ne pas pondérer l'erreur MAE_g selon le nombre N_{60} par ambiance. Par leur durée plus importante, une moyenne pondérée serait plus influencée par les ambiances *rue calme* et *rue bruyante*. Afin de déterminer la combinaison la plus efficace sur l'ensemble des ambiances, le même poids est donc donné à chaque ambiance sonore.

6.2 Erreurs MAE_g obtenues par l'estimateur *baseline*

Dans un premier temps, les erreurs réalisées par l'estimateur *baseline* sont observées sur l'ensemble du corpus et selon chaque ambiance (Tableau 6.3).

TABLEAU 6.3 – Erreurs moyennes MAE_g et MAE_{60} en dB pour l'estimateur *baseline* pour le corpus d'évaluation *SOUR*, en gras-rouge la plus faible erreur MAE_g, en gras-noir les erreurs MAE_{60} les plus faibles.

f_c (Hz)	MAE_g	$MAE_{60,Pa}$	$MAE_{60,Ca}$	$MAE_{60,Br}$	$MAE_{60,TrBr}$
100	2,89 (± 0,56)	**2,48 (± 1,85)**	3,71 (± 2,63)	2,66 (± 1,20)	2,70 (± 0,71)
500	**1,99 (± 1,37)**	3,87 (± 5,08)	**2,14 (± 2,25)**	1,03 (± 0,53)	0,93 (± 0,38)
1k	2,44 (± 2,57)	6,07 (± 5,26)	2,46 (± 2,81)	**0,64 (± 0,64)**	0,59 (± 0,37)
2k	2,99 (± 3,28)	7,60 (± 5,28)	3,01 (± 2,93)	0,95 (± 0,97)	**0,39 (± 0,54)**
5k	3,48 (± 3,84)	8,89 (± 5,22)	3,48 (± 3,12)	1,13 (± 1,10)	0,42 (± 0,64)
10k	3,59 (± 3,93)	9,11 (± 5,29)	3,64 (± 3,19)	1,17 (± 1,14)	0,43 (± 0,66)
20k	3,62 (± 3,93)	9,14 (± 5,31)	3,68 (± 3,20)	1,21 (± 1,14)	0,44 (± 0,67)

De même que dans le chapitre précédent, c'est le filtre PB à la fréquence de coupure f_c = 500 Hz qui génère l'erreur MAE_g la plus faible (MAE_g = 2,03 (± 1,43) dB) en réalisant un compromis entre l'énergie rejetée dans les ambiances *Parc* et *Rue calme* et l'énergie conservée dans les deux autres ambiances. En détaillant les erreurs MAE_{60} selon chaque ambiance, on retrouve un comportement similaire à celui observé pour le corpus *ambiance* : plus la contribution de la source *trafic* est présente, plus la fréquence de coupure nécessite d'être augmentée. On remarque enfin que l'erreur MAE_g à 20 kHz excède 3 dB, valeur qui correspond à la marge d'incertitude acceptée des niveaux de bruits. Cette erreur démontre donc bien que, sans le

traitement des mesures faites en ville, la construction de cartes de bruit trafic serait imparfaite.

6.3 Erreurs MAE_g obtenues par l'estimateur NMF

Les erreurs générées par l'estimateur NMF selon les multiples associations des modalités des facteurs expérimentaux sont maintenant détaillées. Devant le nombre important de résultats, on ne détaille dans le Tableau 6.4 que les résultats les plus performants selon chaque méthode (NMF SUP, SEM et IS) et les 3 valeurs de β.

TABLEAU 6.4 – Erreurs MAE_{60} en dB les plus faibles pour les combinaisons optimales de modalités des estimateurs pour le corpus d'évaluation *SOUR*.

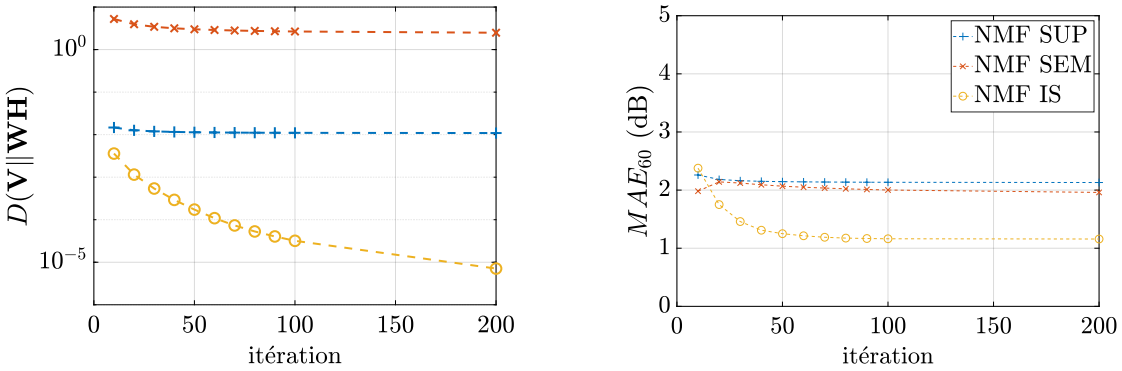
méthode	f_c (kHz)	β	w_t	K	t_h	MAE_{60} (dB)
filtre PB	20	-	-	-	-	3,62 (± 3,93)
	0,5	-	-	-	-	1,99 (± 1,37)
NMF SUP	-	0	0,5	200	-	3,13 (± 3,33)
	-	1	0,5	200	-	2,67 (± 3,02)
	-	2	0,5	25	-	2,13 (± 2,22)
NMF SEM	-	0	0,5	300	-	2,02 (± 0,68)
	-	1	0,5	300	-	1,93 (± 0,42)
	-	2	1	300	-	2,24 (± 0,89)
NMF IS	-	0	1	25	0,34	1,50 (± 1,16)
	-	1	1	200	0,35	1,31 (± 1,02)
	-	**2**	**1**	**300**	**0,35**	**1,16 (± 0,86)**

La composition du corpus *SOUR* étant différente de celle du corpus *Ambiance*, les combinaisons optimales diffèrent de celles obtenues dans le chapitre précédent dans certains cas. La NMF SUP reste la méthode la moins performante avec des erreurs supérieures à celles de l'estimateur *baseline* à $f_c = 500$ Hz avec des écart-types également plus importants notamment pour l'estimateur basé sur la divergence d'Itakura-Saito ($\beta = 0$). Contrairement au corpus *Ambiance* où chaque approche privilégie un dictionnaire comprenant un faible nombre d'éléments, ici, la NMF SUP avec $\beta \in \{0,\ 1\}$ choisit un nombre plus élevé d'éléments ($K = 200$). Seul la NMF basée sur la distance EUC conserve un faible nombre d'éléments.

Les versions optimales de la NMF SEM sont basées sur le même dictionnaire et privilégient un grand nombre d'éléments dans le dictionnaire ($K = 300$). Leur erreurs ne diffèrent donc que par le choix de β. Ces modalités restent ici cohérentes par rapport au corpus précédent où c'était également un dictionnaire unique composé d'un grand nombre d'éléments, qui offrait les erreurs les plus faibles. Pour autant, si la NMF SEM permet des erreurs plus faibles que la NMF SUP, elle n'améliore les performances de l'estimateur *baseline* que pour $\beta = 1$.

La NMF IS se révèle être l'approche la plus performante avec systématiquement, pour les trois valeurs de β, des erreurs MAE_g inférieures à 1,5 dB. La méthode la plus performante est celle basée sur la distance EUC ($\beta = 2$) avec $K = 300$, $w_t = 1$ s et un seuil $t_h = 0{,}35$ et génère

une erreur MAE_g de 1,16 ($\pm$ 0,86) dB. On constate que pour $\beta = 0$, à l'inverse du corpus *Ambiance*, le nombre d'éléments est ici réduit à 25. Les valeurs des seuils sont également plus faibles pour ce corpus *SOUR* puisque la part du trafic est ici plus importante que dans le corpus *Ambiance*. De plus, les classes de sons dans le corpus *SOUR* appartiennent majoritairement aux classes interférantes *alerte*, *animaux* et *humains*, des classes de sons dont les spectres sont moins similaires à ceux du trafic. En conséquence, la distance D_θ entre les éléments *trafic* et ceux modélisant ces sources est plus faible. La NMF IS avec $\beta = 2$, $K = 300$, $w_t = 1$ s et $t_h = 0{,}35$ est donc l'approche qui génère la plus faible erreur de reconstruction sur l'ensemble du corpus d'évaluation *SOUR*. Les combinaisons optimales de la NMF SUP ($\beta = 2$, $K = 25$ et $w_t = 0{,}5$ s) et de la NMF SEM ($\beta = 1$, $K = 300$, $w_t = 2$) génèrent des erreurs plus importantes et ne sont donc pas des estimateurs adéquats pour cette tâche.



(a) Evolution de la fonction de coût moyenne.

(b) Évolution de l'erreur MAE_g.

FIGURE 6.3 – Comparaison de l'évolution de la fonction de coût et des erreurs MAE_g moyennes selon les 3 NMF optimales retenues.

Les erreurs de reconstruction $D(\mathbf{V}\|\mathbf{WH})$ ainsi que l'évolution de l'erreur MAE_g en fonction des itérations sont présentées en Figure 6.3. La fonction de coût est naturellement meilleure pour la NMF IS puisque les matrices $\mathbf{W}$ et $\mathbf{H}$ sont mises à jour. Si la convergence n'est pas atteinte au bout de 200 itération, l'erreur MAE_{60} est pour autant constante. Les mises à jour suivantes modélisent alors les autres sources sonores et non plus le trafic. Entre la NMF SUP et SEM, l'erreur de reconstruction est rapidement stable après 100 itérations. Si la NMF SEM possèdent une erreur MAE_g plus faible que la NMF SUP, son erreur de reconstruction reste toutefois supérieure à celle de la NMF SUP.

6.4 Erreurs MAE_{60} par ambiance sonore

L'erreur MAE_{60} est ensuite détaillée dans le Tableau 6.5 selon les combinaisons optimales de chaque NMF :

— NMF SUP, $\beta = 2$, $w_t = 0{,}5$ s, $K = 25$,

— NMF SEM, $\beta = 1$, $w_t = 0{,}5$, $K = 300$,

— NMF IS, $\beta = 2$, $w_t = 1$, $K = 300$, $t_h = 0{,}35$.

Les erreurs générées par l'estimateur *baseline* à $f_c = 500$ Hz et $f_c = 20$ kHz sont également ajoutées. Le temps de calcul mis par l'estimateur pour extraire la composante *trafic* d'un spectrogramme $\mathbf{V}$ d'une durée de 1 minute est ajouté. Dans le cas de la NMF, c'est le temps mis par la méthode pour réaliser les 200 itérations. Les calculs ont été mené sur un ordinateur équipé d'un processeur Intel Core i7 (64 bits) de 2.40 GHz.

TABLEAU 6.5 – Erreurs MAE_{60} en dB les plus faibles selon les estimateurs NMF pour chaque méthode dans sa combinaison optimale de modalités avec les estimateurs *baseline* à 500 Hz et 20 kHz. En gras-rouge, les erreurs globales les plus faibles, en gras-noir, les erreurs de la NMF inférieures à l'estimateur *baseline* $f_c = 500$ Hz.

méthode	**Parc**	**Rue Calme**	**Rue Bruyante**	**Rue très bruyante**	**temps de calcul (s/minute de signal)**
filtre PB, $f_c = 20$ kHz	9,14 (± 5,31)	3,68 (± 3,20)	1,21 (± 1,14)	0,44 (± 0,67)	0,03 (± 0,01)
filtre PB, $f_c = 500$ Hz	3,87 (± 5,08)	2,14 (± 2,25)	1,03 (± 0,53)	0,93 (± 0,38)	0,03 (± 0,01)
NMF SUP	5,31 (± 5,06)	2,02 (± 2,22)	**0,62 (± 0,56)**	**0,57 (± 0,33)**	0,09 (± 0,03)
NMF SEM	**2,50 (± 2,71)**	2,36 (± 2,42)	1,62 (± 0,87)	1,42 (± 0,60)	1,77 (± 0,13)
NMF IS	**2,13 (± 3,84)**	**1,62 (± 1,85)**	**0,57 (± 0,54)**	**0,32 (± 0,20)**	2,15 (± 0,31)

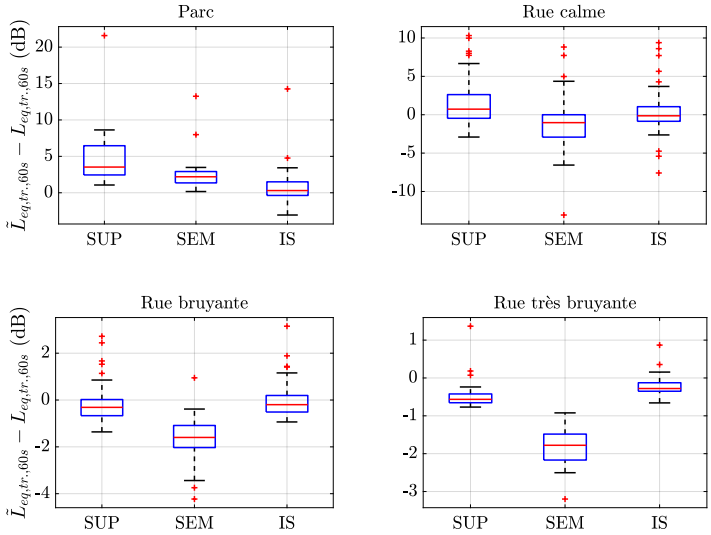


FIGURE 6.4 – Distribution des erreurs relatives selon les 3 NMF optimales retenues et les 4 ambiances sonores.

La Figure 6.4 résume la distribution des erreurs relatives ($\tilde{L}_{eq,tr.,60s} - L_{eq,tr.,60s}$) selon chaque NMF pour les 4 ambiances sonores. Dans le chapitre précédent, la NMF IS était, sur l'intégralité du corpus, la méthode générant l'erreur moyenne la plus faible mais elle ne permettait pas d'obtenir les plus faibles erreurs selon chaque TIR. La NMF SEM optimale générait alors les

erreurs les plus faibles dans les TIR négatifs et la NMF SUP supplantait les autres méthodes pour les TIR positifs. Sur le corpus $SOUR$, la NMF IS se révèle être la méthode obtenant les plus faibles erreurs MAE sur l'ensemble des ambiances sonores ainsi que les erreurs relatives les plus centrées sur zéro. Elle parvient même à être plus performante que le filtre à 20 kHz (équivalent à la mixture globale du signal) dans l'ambiance *Rue très bruyante*. On relève toutefois que dans l'ambiance *Parc*, malgré une erreur MAE_{60} plus faible, l'écart type reste élevé ($MAE_{60} = 2{,}13$ ($\pm$ 3,84) dB) tout en relevant la présence d'outliers qui augmentent alors celle-ci. Le comportement de la NMF SUP et SEM est alors similaire à celui obtenu pour le corpus *Ambiance* : dans les ambiances où le trafic est peu présent, elles sur-estiment le niveau sonore *trafic* puis au fur et à mesure que celui-ci devient prédominant, ces méthodes vont de plus en plus sous-estimer sa valeur, notamment pour la NMF SEM.

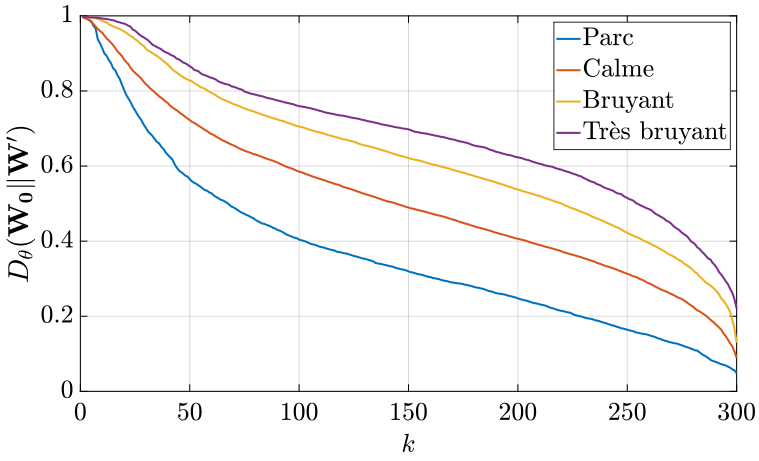


FIGURE 6.5 – Distances $D_\theta(\mathbf{W_0}\|\mathbf{W'})$ moyennes par ambiance sonore triées par ordre décroissant dans le cas de la NMF IS optimale ($\beta = 2$, $K = 300$, $w_t = 1$).

Le comportement de la NMF IS peut s'illustrer à travers les évolutions de la distance D_θ moyenne pour chaque ambiance sonore (Figure 6.5). En fixant un seuil sur l'ensemble des ambiances sonores, le nombre d'éléments inclus dans $\mathbf{W}_{trafic}$ est variable. Dans le cas de l'ambiance *Rue très bruyant*, le nombre d'éléments moyen est de 286 soit quasiment l'intégralité du dictionnaire $\mathbf{W'}$. Ce nombre est justifié par le fait que le trafic est la source sonore dominante dans ces scènes : $\mathbf{W'}$ dévie donc moins de $\mathbf{W_0}$. En utilisant un seuil bas, on conserve suffisamment d'éléments dans $\mathbf{W}_{trafic}$ pour bien modéliser la source *trafic*. Dans le cas de l'ambiance *Parc*, étant composée de peu d'éléments appartenant à la classe de son *trafic*, la distance D_θ dévie beaucoup plus réduisant le nombre d'éléments de $\mathbf{W'}$ inclus dans $\mathbf{W}_{trafic}$ à 130.

Les deux autres NMF sont moins performantes. Le comportement de la NMF SUP reste similaire à celui observé dans le chapitre précédent : elle échoue à améliorer les erreurs de l'estimateur *baseline* dans les ambiances où le trafic est peu présent mais offre de meilleures performances lorsque le trafic devient la source sonore principale grâce à son dictionnaire composé de spectres *trafic*. La NMF SEM est finalement, sur ce corpus et sur les 3 NMF comparées, celle qui offre les

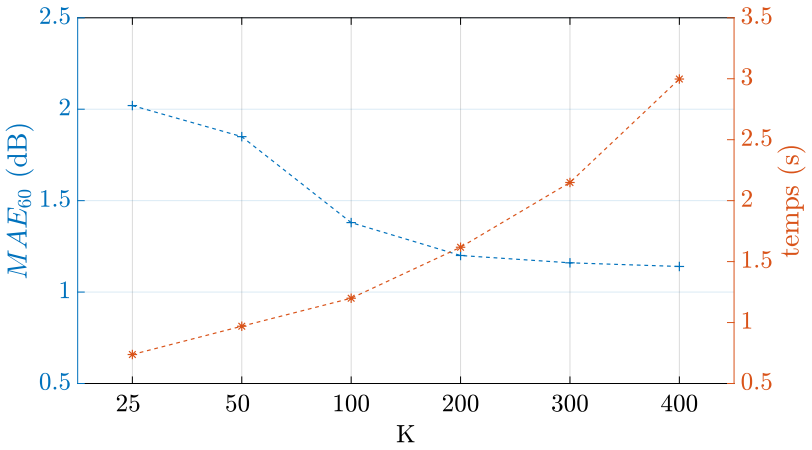


FIGURE 6.6 – Évolution de l'erreur MAE_g et du temps de calcul selon la taille des matrices K dans le cas de la NMF IS optimale (pour la figure, le nombre d'éléments dans K a été étendu à 400).

résultats les moins bons. Si l'erreur obtenue dans l'ambiance *Parc* reste limitée, avec l'augmentation de la présence de trafic, les erreurs ne diminuent pas suffisamment. Cette décroissance limitée a été observée également dans le chapitre précédent et provient de la partie mobile du dictionnaire $\mathbf{W_r}$ qui, dans ces mises à jour, est libre d'intégrer des composantes du trafic routier notamment lorsque le TIR était positif. Ce corpus étant, pour une plus grande part de scènes, dans ce cas là, cette méthode est alors mise en défaut.

Enfin, puisque l'objectif est d'estimer des niveaux sonores du trafic notamment à travers des mesures réalisées par des réseaux de capteurs fixes, le temps de calcul nécessaire à l'estimation de ce niveau est relevé. La dernière colonne du Tableau 6.5 exprime le temps nécessaire à l'estimateur pour extraire la composante *trafic* d'un signal audio de 1 minute. Dans le cas du filtre, ce temps est quasiment instantané puisque l'opération est simple et n'est appliquée qu'une seule fois sur le signal. Pour la NMF, ce temps est plus important puisqu'une minimisation de la distance entre $\mathbf{V}$ et $\mathbf{WH}$ nécessite de réaliser 200 itérations et de mettre à jour une ou deux matrices. C'est en toute logique la NMF SUP qui est la plus rapide car une seule matrice est mise à jour dont le rang est faible ($K = 25$). La NMF IS est, quant à elle, la plus longue puisque les deux matrices $\mathbf{W}$ et $\mathbf{H}$ sont mises à jour et sont de tailles importantes ($K = 300$). Cette méthode nécessite alors plus de 2 secondes pour traiter 1 minute de signal audio. L'influence du nombre d'éléments K dans $\mathbf{W_0}$ sur l'erreur MAE_g ainsi que le temps de calcul nécessaire à la NMF IS pour traiter 1 minute de signal sont résumées en Figure 6.6. Si un plus grand nombre d'éléments dans le dictionnaire implique une description plus fine de la source *trafic*, il implique aussi une augmentation du temps de calcul. Les algorithmes de mises à jour étant multiplicatifs (voir équation 3.37), le nombre d'opérations réalisées, et donc le temps de calcul, augmente nécessairement avec la taille des matrices. Si entre $K = 25$ et $K = 200$, l'erreur chute significativement, à partir de $K = 300$, la variation d'erreur devient faible au regard de la variation du temps de calcul.

La durée du temps de calcul pour $K = 300$ reste cependant faible au regard des utilisations qui peuvent être faites de cette outil : en effet pour la réalisation de cartes de bruit ou la prédiction du bruit de trafic en milieu urbain, connaitre avec une grande précision, c'est-à-dire à la seconde par exemple, le niveau du trafic, n'a pas d'intérêt puisque les environnements sont habituellements caractérisés à des échelles temporelles grandes. [Brocolini *et al.*, 2013b] montre par exemple que 15 minutes sont nécessaires pour capturer les effets perceptifs d'un ESU. Dans le cas de la cartographie, leurs mises à jour dynamiques sont souvent réalisées toutes les heures [Bellucci et Zambon, 2017] ou les quarts d'heure [Wei *et al.*, 2016]. De plus, dans le cas où la mesure a pour but d'améliorer les cartes de bruits via des méthodes d'assimilations de données, ces opérations génèrent un temps de calcul plus élevé qui rend donc la durée de l'estimation du niveau sonore *trafic* par la NMF IS satisfaisante et adéquate.

6.5 Comparaison des niveaux sonores $L_{eq,tr.,1s}$ pour plusieurs scènes sonores.

L'étude des scènes sonores urbaines et l'estimation de leurs niveaux sonores du trafic ont été réalisées sur un temps d'intégration de 1 minute pour le corpus *SOUR*. Si la NMF IS optimale offre la meilleure estimation du $L_{eq,tr.,60s}$ sur l'intégralité du corpus, il est intéressant de visualiser la qualité de la reconstruction du signal *trafic* à une échelle plus réduite. En conséquence, l'évolution du niveau sonore $\tilde{L}_{eq,tr.,1s}$ de 4 extraits de 1 minute issus de chaque ambiance sonore est représentée. Ces extraits sont choisis pour être représentatifs du comportement général observé sur les autres scènes. Ce niveau sonore est comparé à celui exact, $L_{eq,tr.,1s}$ ainsi qu'à celui obtenu par l'estimateur *baseline* à $f_c = 500$ Hz. L'évolution du niveau sonore de la classe de son *interférante* ($L_{eq,int.}$) est également ajoutée. Dans un premier temps, les 60 premières secondes de la scène 5 de l'ambiance *Parc* qui correspond à la réplication de l'enregistrement 2-WE-03 (voir la correspondance des noms des scènes enregistrées et répliquées dans les Tableaux en annexe E) sont observées en Figure 6.7.

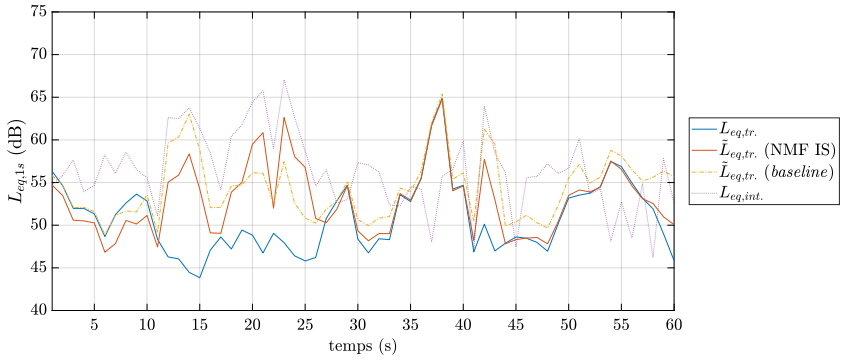


FIGURE 6.7 – Évolution du niveau sonore équivalent $L_{eq,1s}$ durant 60 secondes de la scène *parc-05*.

Cet extrait est notamment composé de deux bruits de fond (*oiseaux*, *trafic routier*) et est

dominé par la classe de son interférante. L'erreur MAE_{60} de l'estimateur NMF est de 1,85 dB alors que celle de l'estimateur *baseline* est de 2,90 dB. On constate que cette erreur, pour les deux méthodes, est due à la classe *interférante*, modélisée par les estimateurs comme un signal *trafic*. L'utilisation de la NMF permet de limiter cette confusion, notamment dans les intervalles [10, 15] secondes et [41, 44] secondes qui correspondent dans les deux cas à l'émergence d'un bruit de rue (barrière). L'estimation du niveaux sonore par la *baseline* suit alors beaucoup plus la classe *interférante* que la NMF IS. Dans l'intervalle [17, 26], la NMF confond plus la classe *interférante* et la classe *trafic* que le filtre à 500 Hz. Cette confusion est réalisée par des bruits de pas.

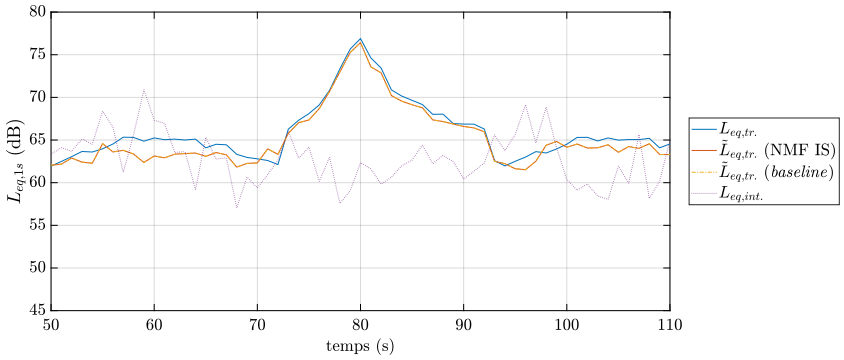


FIGURE 6.8 – Évolution du niveau sonore équivalent $L_{eq,1s}$ durant 60 secondes de la scène *Rue calme-07*.

La seconde scène observée en Figure 6.8 équivaut à un extrait de 1 minute de la 7e scène de l'ambiance *Rue calme* (respectivement la réplication de l'enregistrement 1-EW-17) entre la 50e et la 110e seconde. Elle se compose d'un bruit de fond *oiseaux* et *trafic* et d'évènements *voix* ainsi que d'un passage d'une voiture entre la 70e et la 93e seconde. On observe que les deux estimateurs arrivent à bien estimer l'évolution du $L_{eq,tr.,1s}$ du passage du véhicule. En dehors de cet évènement, les deux estimateurs sont moins soumis à la classe *interférante*. L'erreur, dans cet extrait, est alors due à une sous-estimation du niveau sonore *trafic* hors du passage de la voiture, qui est plus important pour la baseline que pour la NMF IS.

Les Figures 6.9 et 6.10 résument des extraits des scènes *Rue bruyante-07* et *Rue très bruyante-05* (respectivement les répliques des enregistrements 01-EW-06 et 02-EW-16) où le trafic est la source sonore prédominante. L'extrait sonore de la Figure 6.9 est composé d'un bruit de fond de trafic ainsi que de voix et d'oiseaux en continu. Les évènements sonores appartiennent principalement à la classe *trafic*. Malgré cela, si les passages des véhicules émergents entre la 10e et la 40e seconde sont correctement modélisés par les deux estimateurs, la NMF IS arrive ensuite beaucoup mieux à reconstruire le signal *trafic* jusqu'à la fin de l'extrait que la baseline. Dans le cas de l'extrait sonore de la Figure 6.10, la classe de son *interférante* composée de voix est très faible par rapport à la classe de son *trafic*. Sur l'intervalle [33, 72] secondes, les multiples passages de voitures sont correctement estimés par la NMF IS. Si l'estimateur *baseline* suit également cette évolution, il en sous-estime toutefois les niveaux sonores.

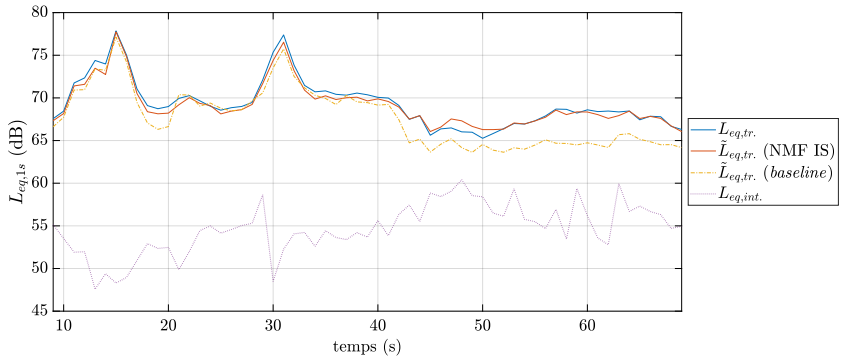


FIGURE 6.9 – Évolution du niveau sonore équivalent $L_{eq,1s}$ durant 60 secondes de la scène *Rue bruyante-03*.

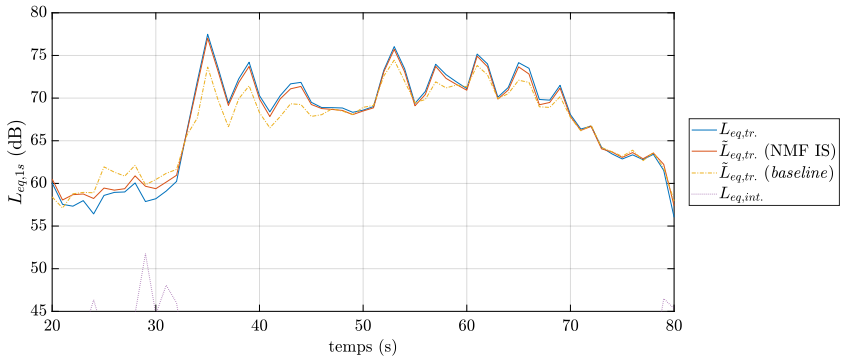


FIGURE 6.10 – Évolution du niveau sonore équivalent $L_{eq,1s}$ durant 60 secondes de la scène *Rue très bruyante-05*.

En conclusion, les erreurs produites par les estimateurs, dans les ambiances *Parc* et *Rue calme*, sont le plus souvent dues à la prise en compte de la composante *interférante* dans le signal *trafic*. Cette confusion reste plus limitée par la NMF IS. Pour les ambiances *Rue bruyante* et *Rue très bruyante*, l'erreur de l'estimateur baseline est notamment due à une sous-estimation du niveau sonore en raison du rejet d'une partie de l'énergie sonore par le filtre passe-bas.

6.6 Pistes d'amélioration

À partir de ces résultats, plusieurs pistes peuvent être envisagées pour améliorer l'estimation du niveau sonore du trafic. Parmi elles, l'ajout de la contrainte de régularité temporelle sur les activateurs est testé.

6.6.1 Contrainte de régularité temporelle

Cette contrainte, présentée dans la partie 3.8.2, consiste à forcer les activateurs temporels de la matrice $\mathbf{H}$ à adopter des variations plus lentes entre les trames temporelles. Son utilisation est justifiée ici par le comportement de la source d'intérêt : le passage d'une voiture est un évènement

qui a une durée de plusieurs secondes. En forçant les activateurs à adopter des variations plus lentes, on serait susceptible de pouvoir mieux modéliser cette source. Cette contrainte se traduit par l'ajout d'un second terme à la fonction de coût [Virtanen, 2007],

$$\min D\left(\mathbf{V}\|\mathbf{WH}\right) + \alpha_t C_t(\mathbf{H}) \quad \text{avec} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{6.4}$$

avec la contrainte de régularité temporelle (ou *smoothness*) $C_t(\mathbf{H})$ présentée dans la partie 3.8.2 par l'équation 3.57 et rappelée :

$$C_t(\mathbf{H}) = \sum_{n=1}^{K} \sum_{n=2}^{N} \left(h_{kn} - h_{k(n-1)}\right)^2, \tag{6.5}$$

et la pondération α_t qui modifie l'importance de cette contrainte dans l'erreur de reconstruction. Cette pondération peut être variable selon K ($\alpha_{t,k}$) et sera alors insérée dans la somme. La pondération α_t est définie pour plusieurs valeurs : $\alpha_t \in \{0,01,\ 0,05,\ 0,1,\ 0,5,\ 1,\ 2,\ 3,\ 5\}$ et s'applique sur les trois versions de la NMF testées pour l'ensemble des combinaisons testées dans la partie 6.3. Dans les cas de la NMF SUP et la NMF IS, cette contrainte s'applique sur l'ensemble des éléments de la matrice $\mathbf{H}$ puisqu'elle est composée entièrement de spectres *trafic*. Leurs mises à jour suivent donc l'équation 3.58. Dans le cas de la NMF SEM, le dictionnaire se décompose en deux parties (une partie fixe $\mathbf{W_s}$ composée de spectres *trafic* et une partie libre $\mathbf{W_r}$). Pour favoriser les variations lentes des activateurs temporels des spectres *trafic*, cette contrainte est seulement posée pour $\mathbf{H_s}$. La mise à jour de la matrice suit donc l'équation 3.58 alors que celle de $\mathbf{H_r}$ reste la même (équation 3.37b). Le nombre d'itérations est étendu à 400 afin d'obtenir une convergence satisfaisante des fonctions de coût. Les erreurs MAE_g minimales obtenues avec la NMF SUP, SEM et IS pour chaque valeur de β sont résumées dans le Tableau 6.6. En plus de cela, les erreurs minimales obtenues par chaque méthode dans la partie 6.3 sont également adjointes en tant que références.

L'ajout de la contrainte de *smoothness* a un faible impact sur la NMF SUP pour $\beta \in \{1,\ 2\}$. Dans le cas de la divergence KL, cette contrainte génère un changement dans la composition du dictionnaire avec un faible nombre d'éléments ($K = 25$). Dans le cas de la distance EUC, l'influence de la contrainte est trop faible pour pouvoir être considérée.

Dans le cas de la NMF SEM, l'influence de la pondération est positive pour la divergence KL et la divergence IS. Dans ce dernier cas, pour un dictionnaire de même taille (avec une fenêtre w_t toutefois différente), la diminution est importante avec une erreur $MAE_g = 1,53$ dB alors qu'elle était de 2,02 dB sans contrainte. Dans la Figure 6.11, les erreurs MAE_{60} de la NMF SEM pour la méthode optimale sans pondération ($\beta = 1$, $K = 300$, $w_t = 0{,}5$ s) et pour $\beta = 0$, $K = 50$, $w_t = all$ sans et avec pondération, sont comparées. Les erreurs MAE_{60} basées sur la NMF SEM avec la divergence IS évoluent différemment par rapport aux autres NMF vues jusqu'à présent. Sans pondération, elle génère dans l'ambiance *Parc* une erreur beaucoup plus faible que dans les autres ambiances sonores ($MAE_g = 1{,}09$ dB). Mais avec l'augmentation de la présence du trafic, cette méthode échoue à obtenir des estimations du niveau sonore du trafic correctes.

TABLEAU 6.6 – Erreurs MAE_{60} les plus faibles pour les combinaisons optimales de modalités des estimateurs pour le corpus d'évaluation *SOUR* en présence d'une pondération de régularité temporelle.

méthode	f_c (kHz)	β	w_t	K	t_h	α_t	MAE_{60} (dB)
filtre PB	20	-	-	-	-	-	3,62 (± 3,93)
	0,5	-	-	-	-	-	1,99 (± 1,37)
NMF SUP	-	2	0,5	25	-	0	2,13 (± 2,22)
	-	0	2	100	-	3	3,54 (± 3,75)
	-	1	2	25	-	3	2,56 (± 2,58)
	-	2	0,5	25	-	0,01	2,12 (± 2,14)
NMF SEM	-	1	0,5	300	-	0	1,93 (± 0,42)
	-	0	2	300	-	0,5	1,53 (± 0,81)
	-	1	0,5	300	-	0,5	1,84 (± 0,34)
	-	2	2	300	-	5e-3	2,19 (± 0,61)
NMF IS	-	2	1	300	0,35	0	1,16 (± 0,86)
	-	0	*all*	50	0.45	0,05	1,50 (± 1,11)
	-	1	1	200	0,36	0,50	1,37 (± 0,76)
	-	2	1	300	0,34	0,01	1,48 (± 0,90)

L'ajout de la contrainte de *smoothness*, si elle détériore les performances dans l'ambiance *Parc*, permet alors une meilleure estimation du signal *trafic* pour les 3 autres ambiances sonores.

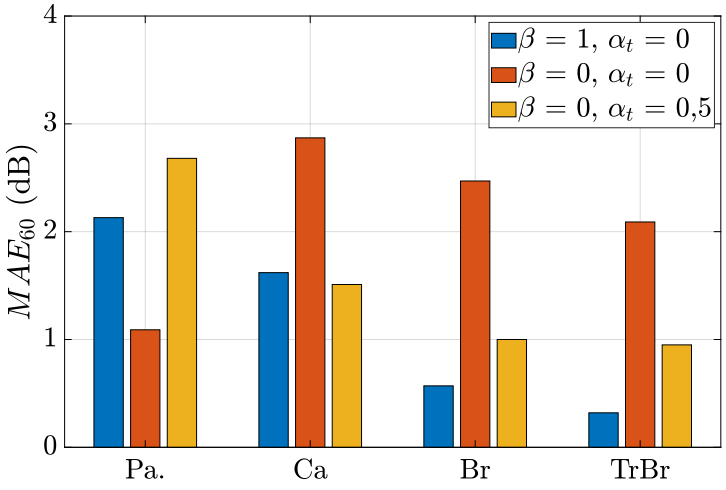


FIGURE 6.11 – Influence de la pondération de régularité temporelle pour la NMF SEM selon chaque ambiance sonore.

L'impact de cette contrainte peut être observé sur le spectre du signal obtenue par le produit des éléments libres $\mathbf{W_r H_r}$ de la NMF SEM avec $\beta = 0$, $K = 50$ et $w_t = all$ dans deux scènes : une extraite de l'ambiance *Parc* (Figure 6.12(a)) et l'autre dans l'ambiance *Rue très bruyante* (Figure 6.12(b)). Pour chacune on ajoute également le spectre du signal interférant. Dans les deux cas, l'ajout de la contrainte a un impact sur la forme du spectre du signal $\mathbf{W_r H_r}$. Pour

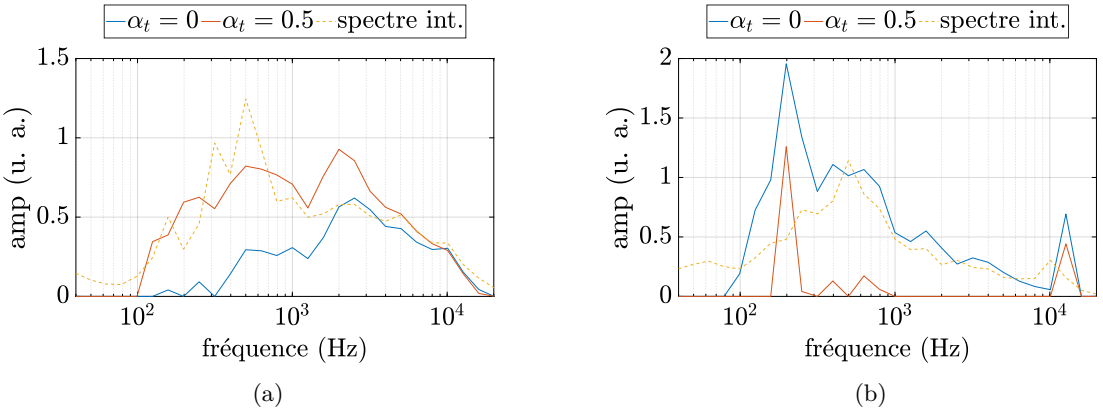
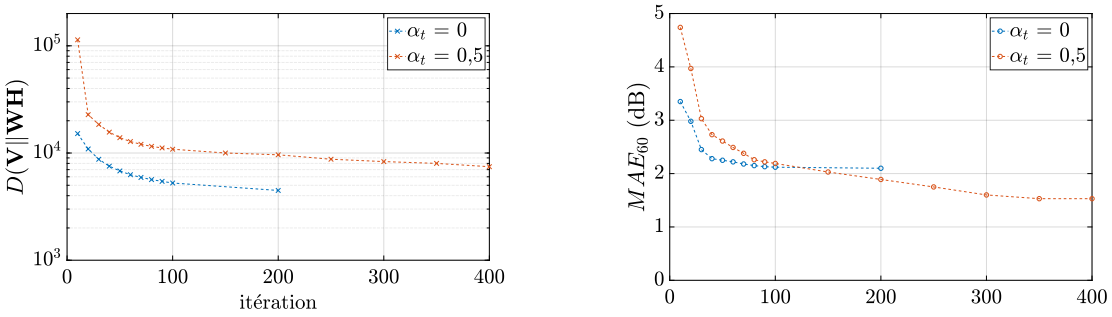


FIGURE 6.12 – Comparaisons des spectres de la partie libre $\mathbf{W_r H_r}$ pour le cas sans pondération ($\alpha_t = 0$) et avec ($\alpha_t = 0,05$) pour la scène 2 de l'ambiance *Parc* (a) et la scène 7 de *Rue très bruyante* (b).

la scène *Parc*, la présence de trafic étant plus réduite et la partie fixe du dictionnaire étant contrainte dans son fonctionnement, la composante *trafic* est moins bien modélisée par la partie fixe $\mathbf{W_s}$. Ainsi, le dictionnaire $\mathbf{W_r}$ vient en complément afin de minimiser la distance entre $\mathbf{V}$ et $\mathbf{WH}$. Cette apport se traduit alors par une plus grande présence de composantes dans les basses fréquences dans $\mathbf{W_r}$ avec la pondération. Cet ajustement détériore donc l'estimation du niveau sonore du trafic. Pour la scène *Rue très bruyante*, l'ajout de la contrainte a un aspect positif : étant la source principale, la contrainte peut alors être perçue comme moins « difficile » ce qui permet ainsi de limiter l'intégration de composantes *trafic* dans $\mathbf{W_r}$ et ainsi de réduire les erreurs MAE_{60}.

Enfin, l'impact de la contrainte sur la distance $D(\mathbf{V}\|\mathbf{WH})$ et sur l'évolution de l'erreur sont observées. Cette contrainte augmente les valeurs de la fonction de coût qui converge également plus lentement. Après 400 itérations, celle-ci reste supérieure à celle sans pondération obtenue après 200 itérations. Sous la contrainte de régularité temporelle, la reconstruction globale du signal semble donc moins bonne. Toutefois, au regard de l'erreur MAE_{60}, la reconstruction de la composante *trafic* est de meilleure qualité. En conclusion, la NMF SEM permet une plus grande capacité d'adaptation aux différentes scènes sonores grâce aux degrés de liberté apportés par son dictionnaire mobile $\mathbf{W_r}$ par rapport à la NMF SUP. Mais cette méthode gagne à être contrainte afin de limiter cette liberté pour mieux focaliser sur la source sonore *trafic*. Si dans les scènes où le trafic est peu présent, cette contrainte peut se relever être un désavantage, dans le cas où le trafic devient prépondérant, celle-ci permet non pas de mieux modéliser la composante interférante dans $\mathbf{W_r}$ mais de limiter la présence de celle du trafic.

Finalement, c'est pour la NMF IS que la pondération a le moins d'impact puisque les erreurs minimales obtenues sont supérieures aux cas sans pondération quelle que soit la valeur de β. À titre de comparaison, les erreurs MAE_{60} sont résumées dans la Figure 6.13 pour les combinaisons optimales sans pondération ($\beta = 2$, $w_t = 1$, $K = 300$, $t_h = 0{,}35$) et pour celle avec la pondération ($\beta = 1$, $w_t = 1$, $K = 200$, $t_h = 0{,}36$, $\alpha_t = 0{,}50$). Si pour les ambiances *Parc* et *Rue calme*, la



(a) Evolution de la fonction de coût moyenne.

(b) Évolution de l'erreur MAE_g.

FIGURE 6.13 – Impact de la contrainte de régularité temporelle sur la fonction de coût et les erreurs MAE_g moyennes pour la NMF SEM avec $\beta = 0$, $w_t = all$ s et $K = 50$.

contrainte a peu d'influence, celles-ci dégrade les performances de la NMF lorsque le trafic devient la source principale. La contrainte imposée à la NMF IS ne lui permet alors pas de modéliser au mieux la source sonore.

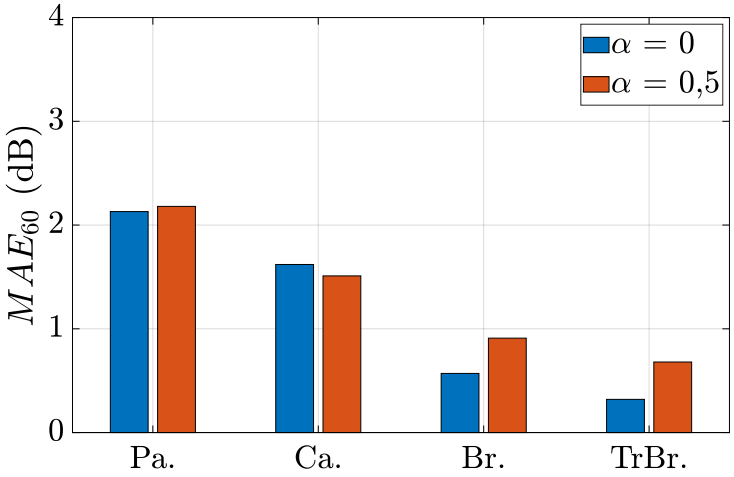


FIGURE 6.14 – Influence de la contrainte de régularité temporelle sur la NMF IS pour leur combinaison optimale de modalités ($\beta = 1$, $w_t = 1$, $K = 200$, $t_h = 0{,}36$) pour $\alpha_t = 0$ et $\alpha_t = 0{,}5$.

6.6.2 Optimisation par les environnements sonores

La NMF IS optimale trouvée dans la partie 6.3 ($\beta = 2$, $K = 300$ et $w_t = 1$ s) est définie selon un seuil fixe sur l'ensemble du corpus et des ambiances sonores. Toutefois, il peut être possible de diminuer ces erreurs en adaptant le seuil t_h à chaque environnement sonore. Cette adaptation peut être facilement mise en place pour les réseaux de capteurs fixes par l'estimation de l'environnement urbain aux alentours du microphone. Une fois celui-ci déterminé le seuil correspondant à l'ambiance sonore est choisi. En considérant la NMF IS optimale avec $\beta = 2$, K

= 300 et $w_t = 1$ s, les seuils optimaux $t_{h,opt.}$ et les erreurs MAE_{60} correspondantes par ambiance sont résumés dans le Tableau 6.7.

TABLEAU 6.7 – Erreurs MAE_{60} minimales selon le seuil optimal $t_{h,opt}$ par ambiance sonore, en gras les erreurs minimales.

ambiance	**Parc**	**Rue Calme**	**Rue Bruyante**	**Rue très bruyante**	MAE_{60} (dB)
seuil fixe t_h	0,35	0,35	0,35	0,35	0,35
erreur MAE_{60} (dB)	2,13 (± 3,84)	1,62 (± 1,85)	0,57 (± 0,54)	0,32 (± 0,20)	1,16 (± 0,86)
seuil optimal $t_{h,opt.}$	0,38	0,35	0,33	0,31	-
erreur MAE_{60} (dB)	**2,03 (± 3,47)**	**1,62 (± 1,85)**	**0,56 (± 0,67)**	**0,28 (± 0,31)**	**1,12 (± 0,83)**

La détermination de chaque seuil optimal permet logiquement de diminuer les erreurs MAE_{60}, cette optimisation a toutefois un impact très limité sur l'erreur globale MAE_g. Dans le cas de *Rue calme*, le seuil optimal $t_{h,opt.}$ correspond même au seuil fixe t_h. Cette limitation est due à l'impact des seuils sur l'évolution de l'erreur. La Figure 6.15 résume l'erreur MAE_{60} par ambiance sonore en fonction de la valeur du seuil t_h.

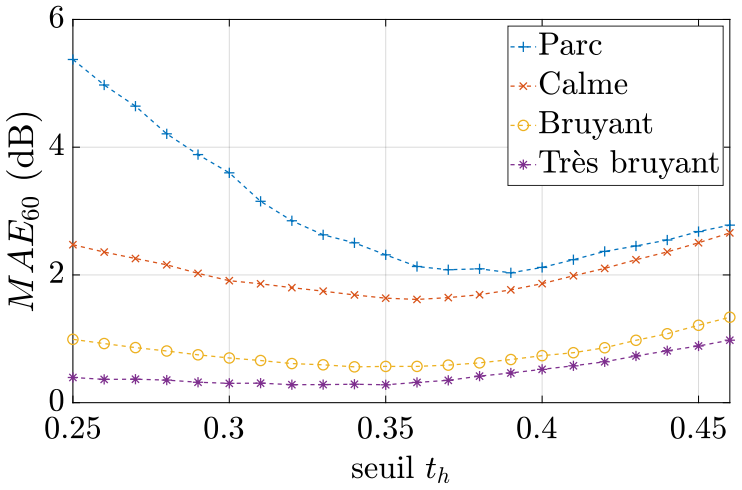


FIGURE 6.15 – Influence de la valeur seuil t_h sur les erreurs MAE_{60} selon l'ambiance sonore.

Dans le cas des ambiances *Rue bruyante* et *Rue très bruyante*, les erreurs sont peu sensibles aux variations du seuil. Le dictionnaire est alors composé d'un grand nombre d'éléments *trafic* (286 éléments), la variation du seuil modifie ce nombre de 1 ou 2 éléments, ce qui est relativement faible par rapport à la taille du dictionnaire $\mathbf{W}_{trafic}$. De plus dans le cas de l'ambiance *Rue très bruyante*, la diminution du seuil jusqu'à $t_h = 0{,}31$, et donc l'augmentation du nombre d'éléments dans $\mathbf{W}_{trafic}$, permet de réduire l'erreur MAE_{60}. Celle-ci devient par la suite quasi constante malgré la possibilité d'inclure dans $\mathbf{W}_{trafic}$ des éléments de la classe *interférante*. Comme l'impact de cette classe est alors très faible dans ces scènes sonores et que le nombre d'éléments *trafic* est beaucoup plus élevé, leur prise en compte dans $\mathbf{W}_{trafic}$ aura une influence très faible. Pour l'ambiance *Parc*, le choix de $t_{h,opt}$ est plus sensible au choix de la valeur du seuil.

Cette sensibilité est d'autant plus forte que les valeurs seuils sont faibles en raison, ici, de la prédominance de la classe *interférante*. Son augmentation permet alors de limiter leur présence dans $\mathbf{W}_{trafic}$.

Ainsi, même si le choix d'utiliser des valeurs seuils optimales à l'ambiance sonore parait intéressant, la faible sensibilité de l'erreur MAE_{60} pour les ambiances sonores *Rue calme*, *Rue bruyante* et *Rue très bruyante* ne permet pas d'avoir une diminution significative des erreurs. Cette amélioration est surtout présente pour l'ambiance *Parc* qui permet de limiter la présence d'éléments de la classe *interférante* dans $\mathbf{W}_{trafic}$. En annexe C est présenté un travail menant à l'estimation d'un seuil optimal à l'aide d'indicateurs de niveau sonore. Le seuil optimal est ainsi non pas déterminé *a priori* mais à partir de la valeur d'un indicateur corrélé à l'évolution des seuils optimaux $t_{h,opt}$ permettant ainsi de s'adapter à des environnements variables. Cette piste peut être utile dans le cas de mesures où l'ambiance sonore n'est pas définie au préalable ou si l'environnement sonore évolue au cours du temps. Elle ne permet pas pour autant d'améliorer suffisamment les estimations des niveaux sonores *trafic* pour être considérée ici.

Les différentes pistes abordées en vue d'améliorer l'estimation du niveau sonore du trafic n'ont pas permis d'apporter de changements significatifs dans le cas de la NMF IS. L'utilisation de contrainte dans la NMF a détérioré les estimations du niveau sonore du trafic. Cette approche semble donc plus performante lorsque les mises à jour de ces matrices est libre. À l'inverse, l'apport de contrainte sur la NMF SEM a eu des impacts positifs sur plusieurs associations de modalités. L'apport de contrainte a ainsi permis de limiter la liberté de la méthode et ainsi de mieux rester focalisée sur la source *trafic*. Enfin l'adaptation des seuils aux environnements sonores ne diminue pas les erreurs d'estimation suffisamment pour être intéressante cela en raison de la faible sensibilité des erreurs aux valeurs des seuils pour les ambiances *Rue calme*, *Rue bruyante* et *Rue très bruyante*.

6.7 Conclusion du chapitre

L'utilisation de la NMF sur le corpus d'évaluation *SOUR* a permis d'identifier l'approche de la NMF la plus performante pour estimer le niveau sonore du trafic. Celle-ci est basée sur la NMF initialisée seuillée avec une distance Euclidienne ($\beta = 2$), 300 éléments dans le dictionnaire et une fenêtre temporelle w_t de 1 seconde. Par la mise à jour intégrale de son dictionnaire initial $\mathbf{W_0}$, composé de spectres *trafic* et de leur classement après 200 itérations, cette méthode réussit à s'adapter à l'ensemble des différentes ambiances sonores ce que ne permet pas la NMF supervisée et semi-supervisée. Les essais pour améliorer les performances des NMF par l'ajout de contraintes de régularité temporelle et de parcimonie (voir Annexe D) n'ont pas abouti à une amélioration des performances des estimateurs. Leurs utilisations nécessitent toutefois d'être approfondies par d'autres études. Enfin, l'utilisation de seuils optimaux à chaque ambiance sonore est également une piste possible. Une première étude, résumée en annexe C, permet d'obtenir des erreurs plus faibles à partir d'indicateurs de niveaux, mais nécessite, là encore, d'autres études.

Conclusions générales et perspectives

L'objectif de cette thèse a été de proposer un outil de traitement du signal visant à déterminer la contribution sonore du bruit du trafic routier à partir de mesures acoustiques réalisées en milieu urbain. L'utilisation de mesures acoustiques en ville s'est vu être de plus en plus considérée afin d'apporter une meilleure description des environnements sonores urbains. Ces mesures permettent alors d'en obtenir une description plus globale en prenant en compte l'ensemble des sources présentes. Mais elles offrent aussi la perspective de pouvoir connaitre les différentes contributions de nombreuses sources présentes en ville. Parmi elles, le trafic routier est la source présentant le plus d'intérêt aux vues des nombreuses études et outils dont il fait l'objet. La réalisation de mesures permettrait notamment l'amélioration de la cartographie de bruit de trafic qui se base, à l'heure actuelle uniquement sur des modèles prédictifs d'émission et de propagation sonore. Toutefois, leur utilisation pose plusieurs questions auxquelles ces travaux ont apporté une première réponse : comment isoler la composante trafic des mesures ? Comment évaluer la qualité de cette séparation ? Afin de répondre à ces question, un outil de traitement de signal a été construit basé sur une méthode de séparation de sources : la Factorisation en Matrice Non-négatives (NMF). Cette méthode présente l'intérêt d'être adaptée aux capteurs individuels, qui équipent les réseaux de mesures ou bien encore les smartphones, et de naturellement prendre en compte le recouvrement des sources sonores, phénomène récurrent en ville. Afin d'évaluer ces performances, l'application de la NMF a été testée non pas sur des enregistrements audio mais sur des corpus de scènes sonores simulées afin d'en contrôler l'apparition des classes de sons, leur niveau sonore... Ce choix a notamment permis de connaitre le niveau sonore du trafic exact qui est ainsi comparé au niveau sonore estimé par la NMF. La mise en place d'un protocole expérimental innovant a permis ensuite d'étudier le comportement de différentes approches de la NMF et d'évaluer ces performances.

Bilan de la thèse

Afin de trouver une forme optimale de la NMF, de nombreux aspects algorithmiques et de paramétrisation de cette méthode ont été abordés à travers la composition du dictionnaire (nombre d'éléments, dimension des trames temporelles) basé sur des enregistrements de passages de voitures, le choix de la β divergence ou la forme de la NMF à travers :

— la NMF supervisée [Lee et Seung, 1999, Févotte et Idier, 2011], qui est l'approche la plus

simple, composée d'un dictionnaire fixe $\mathbf{W}$ constitué d'éléments *trafic* et où seul la matrice d'activation $\mathbf{H}$ est mise à jour,

— la NMF semi-supervisée [Lee *et al.*, 2010, Kitamura *et al.*, 2014], qui inclut dans son dictionnaire une partie fixe $\mathbf{W_s}$ composé de spectres sonores du *trafic* et d'une partie libre $\mathbf{W_r}$ mise à jour pouvant intégrer d'autres sources sonores,

— la NMF initialisée seuillée, développée dans le cadre de ces travaux, qui consiste en l'apprentissage supervisé d'un dictionnaire initial, $\mathbf{W_0}$, composé d'éléments reliés à la source *trafic*. Ce dictionnaire est ensuite mis à jour sur la scène testée. Chaque élément du dictionnaire obtenu, $\mathbf{W'}$, est alors comparé à son état initial dans $\mathbf{W_0}$. Par une technique de seuillage dur, les éléments les plus similaires à leurs états initiaux, et donc susceptibles d'être encore liés au trafic, sont alors conservés et permettent de déterminer la composante *trafic* du signal.

L'ensemble de ces méthodes ont été comparés au regard d'une méthode de référence permettant de contextualiser l'intérêt de ces méthodes. Considérant que le trafic est majoritairement composé spectralement de basses fréquences, nous avons adopté une approche heuristique constituée d'un filtre passe-bas de fréquence de coupure $f_c = 500$ Hz.

L'étude de ces méthodes a été réalisé sur un premier corpus, le corpus *Ambiance*. Celui-ci a été construit en mélangeant une composante *interférante* définie selon certaines classes de sons spécifiques avec une composante *trafic* dont le niveau sonore a été calibré à certains niveaux. Ce corpus, artificiel dans sa forme, a permis d'étudier le comportement de la NMF selon la présence du trafic ou d'une classe de son interférante particulière. Les résultats obtenus soulève la difficulté d'obtenir une méthode adaptée à l'ensemble à des différentes classes de son avec une présence de la composante trafic variable. L'utilisation de la NMF SUP, constituée d'un dictionnaire *trafic*, échoue à obtenir des estimations satisfaisantes du niveau sonore du trafic lorsque celui-ci est peu présent en confondant les classes de sons *trafic* et *interférante* mais elle devient performante lorsque le bruit du trafic devient prépondérant. À l'inverse, l'approche basée sur la NMF semi-supervisée, par l'ajout d'éléments libres dans le dictionnaire, se trouve être performante quand le trafic est peu présent en y intégrant la classe de son interférant, laissant le dictionnaire fixe, composé d'éléments *trafic*, modéliser cette source. Mais, elle devient défaillante lorsque cette source devient principale en considérant dans cette partie libre, des composantes liées au trafic routier. Les degrés de libertés ajoutés sont ainsi, suivant la présence du trafic, un atout ou une faiblesse : avec peu de trafic, ces éléments intègrent facilement les composantes de la classe de son *interférante* alors que face à des mixtures sonores composées principalement de la source *trafic*, celle-ci est libre d'y inclure cette composante, diminuant les performances de la méthode. La NMF initialisée seuillée semble alors être la méthode qui offre les meilleures estimations moyennes même si cette méthode se révèle n'être jamais la plus efficace selon les différentes valeurs du TIR (*Traffic Interfering Ratio*, le rapport des niveaux sonores de la composante *trafic* et de la composante *interférante*). Par la mise à jour de son dictionnaire initial $\mathbf{W_0}$ et au contrôle réalisé par le seuillage, cette méthode s'adapte à chacune des scènes et permet de résoudre la question de la généralisation du dictionnaire. De plus, cette approche

permet, contrairement à la NMF SUP et SEM, de modéliser les signaux tels que captés par un microphone en raison de la mise à jour du dictionnaire. Ainsi, l'impact de l'environnement urbain sur la propagation du son peut être considéré et ainsi être plus généralisable que les deux autres méthodes qui contiennent un dictionnaire fixe. La NMF IS montre toutefois des limites face aux classes de sons ayant des allures spectrales similaires comme les évènements mécaniques (bruit de chantier, ventilation) et climatiques (pluie ou orage) où elle ne permet pas de différencier suffisamment ces évènements du trafic. Concernant les évènements climatiques, la présence de capteurs météorologiques dans les systèmes embarqués permettrait d'éviter de prendre en compte les mesures acoustiques en présence de pluie ou d'orage par exemple.

Les tests menés sur le corpus *SOUR* ont permis d'évaluer les performances de la NMF sur un corpus dont les mixtures sonores qui le compose sont assimilables à des enregistrements sonores urbains. Cet aspect « réaliste » provient de la construction de ces scènes qui s'est basé sur l'annotation et l'écoute d'enregistrements audio urbains. Les résultats obtenus confirment alors les conclusions faites : la NMF IS est la méthode qui s'adapte le mieux aux différents ESU avec en moyenne une erreur globale $MAE_g = 1{,}16$ ($\pm$ 0,86) dB pour une distance euclidienne, un nombre d'éléments $K = 300$, une fenêtre temporelle $w_t = 1$ s et un seuil dur $t_h = 0{,}35$. Cette combinaison obtenue peut ainsi être considérée en vue de déterminer le niveau sonore du trafic par des mesures faites en ville grâce à la qualité des scènes sonores simulées. En effet, le test perceptif mené sur une partie des scènes du corpus sur un panel de 50 auditeurs a permis d'évaluer et de valider le critère de « réalisme » de ces scènes simulées permettant de les assimiler à des enregistrements sonores. L'erreur réalisée par la NMF sur le corpus *SOUR* peut alors être celle qui serait faites sur des enregistrements sonores urbains. L'ajout de contraintes (régularité temporelle et de parcimonie) sur la NMF IS n'a pas permis d'améliorer ces résultats. Celles-ci ont surtout montré un impact pour la NMF SEM où appliquées sur la partie fixe ou mobile du dictionnaire, elles ont pu améliorer l'estimation du niveau sonore du trafic sans toutefois surpasser la NMF IS. Enfin, l'utilisation de seuils adaptatifs à chaque ambiance sonore ne permet pas d'améliorer significativement ces estimations.

Perspectives de recherches

En l'état actuel, les différents résultats obtenus permettent donc de privilégier la NMF IS comme l'approche la mieux adaptée aux environnements sonores urbains en vue d'estimer le niveau sonore du trafic. Cette méthode surpasse les approches plus classiques de la NMF en permettant une meilleure adaptation du dictionnaire aux différentes scènes et ambiances sonores. Cette méthode permet alors de considérer les sources sonores présentes mais également l'effet de l'environnement urbain sur leur propagation. De plus, elle pourra être considérée dans des réseaux de capteurs embarqués en vue d'estimer le niveau sonore du trafic. Les travaux réalisés durant cette thèse, qui ont permis d'obtenir ce résultat, ont naturellement abordés d'autres aspects qui n'ont pas pu être approfondis ou étudiés et qui pourraient améliorer les conclusions ou les performances de cette méthode.

Amélioration de la NMF

Les pistes d'amélioration de la NMF, abordées en fin de thèse, nécessiteraient d'être approfondies notamment celles liées à la contrainte de parcimonie. En plus de ces contraintes, la prise en compte des effets de la propagation du son par l'ajout de filtre de propagation peut être une piste d'étude intéressante. Si pour la NMF IS, ce filtre ne semble pas nécessaire, puisque cet aspect est pris en compte par la mise à jour du dictionnaire, il pourrait être considéré pour la NMF SUP et SEM en vue de leur offrir une meilleure adaptabilité aux différentes scènes sonores urbaines. Il peut être envisageable soit de dupliquer plusieurs fois un dictionnaire fixe et de les filtrer par différentes filtres de propagation afin d'obtenir plusieurs formes de dictionnaires soit de contraindre les mises à jour de $\mathbf{W}$ à suivre l'effet d'un filtre de propagation afin de conserver des éléments *trafic*. Ensuite, dans le cadre de la NMF IS, il peut être envisageable de construire cette méthode sous une forme différente qui serait basée sur la NMF SEM : le dictionnaire pourrait être composé d'un dictionnaire initial $\mathbf{W_0}$ auquel serait ajouté quelques éléments libres en plus. L'ensemble du dictionnaire serait alors mis à jour et les éléments *trafic* seraient extraits par le même processus de seuillage. L'ajout de contrainte, notamment de parcimonie, sur ces éléments libres, pourrait aider la focalisation des mises à jour vers la source d'intérêt ou éviter que les deux éléments prennent la forme de spectres *trafic*.

Validation des conclusions sur d'autres corpus de sons

La NMF IS est la méthode qui a montré les meilleures performances pour estimer le niveau sonore du trafic. Cette conclusion gagnerait naturellement à être validée sur d'autres corpus de sons plus conséquents et variés. Le corpus *SOUR* est représentatif d'un environnement sonore particulier, celui d'une ville occidentale, et ne comprend pas certaines sources sonores comme le bruit de fontaines, de passages de train ou de tramway par exemple. De plus, dans la conception de scènes sonores, d'autres classes de sons, relatifs au trafic routier, n'ont pas pu être enregistrées comme le passage de bus ou de deux roues motorisés (moto, scooter). La base de données élémentaire de sons gagnerait à obtenir des enregistrements de passages de ces véhicules de la même qualité que ceux des voitures afin d'apporter une plus grande diversité. L'application de la NMF IS à de nouveaux corpus de scènes sonores urbaines est donc nécessaire.

Création généralisée de scènes sonores urbaines

Un autre aspect lié à la création de scènes sonores n'a également pas pu être étudié durant ces travaux : l'utilisation dans le logiciel *SimScene* des paramètres de hauts niveaux (classes de sons, niveaux sonores, occurrences de chaque classe de son par ambiance) relevés lors des annotations des enregistrements *GRAFIC* (voir Tableau 4.8). Ces paramètres peuvent être utilisés dans le simulateur *SimScene* afin de générer plus aléatoirement des scènes sonores urbaines. L'utilisation de tels paramètres ouvrirait alors la possibilité de pouvoir composer de plus larges corpus de scènes sonores urbaines tout en étant basés sur des données issues d'enregistrements sonores. Leur utilisation pour composer des scènes n'a pas pu être réalisée et validée durant cette thèse.

Afin de consolider les valeurs extraites, il est nécessaire de réaliser de nouvelles annotations sur d'autres enregistrements sonores urbains, de créer des scènes sonores à partir de ces paramètres pour ensuite les soumettre à un test perceptif. Là où le corpus *SOUR* possédait une structure écologiquement valide puisque basée sur des enregistrements, ce corpus serait plus construit à partir de données plus aléatoires. Ces travaux impliquent aussi de compléter et d'enrichir la base de données élémentaires d'autres sources sonores (camion, bus, deux roues motorisés) mais surtout d'enregistrements de voix basés sur le ton de la discussion. En effet, une limite cette base de données est actuellement la composition des voix qui n'est pas suffisante pour atteindre le degré de réalisme souhaité. Celle-ci se compose de mots, de sons. Lors du test perceptif mené, les quelques évènements *voix* évalués par les auditeurs ont eu un effet négatif sur l'évaluation du réalisme. Les voix présentes dans les enregistrements *GRAFIC* sont composées de personnes qui discutent et rient à plusieurs, parlent au téléphone... Aucune base de données libre proposant de tels extraits sonores n'a été trouvée, il serait donc utile d'en créer une afin de satisfaire le réalisme des scènes sonores urbaines simulées. La piste la plus rigoureuse et souhaitable serait d'enregistrer des dialogues écrits et lus par des comédiens dans des studios d'enregistrements et cela dans plusieurs langues. Malgré cela, les corpus d'évaluation *Ambiance* ou *SOUR*, bien que perfectibles, restent, en l'état actuel, de qualité suffisante pour être utilisés par les communautés dédiés à la création d'outils de traitement du signal pour des tâches de détection, de séparation de sources ou de classification de scènes sonores.

Utilisation de l'apprentissage profond et des réseaux de neurones

Une méthode émergente depuis plusieurs années dans la communauté du *machine learning* doit également être testée sur ces corpus : l'apprentissage profond couplé aux réseaux de neurones [Schmidhuber, 2015, LeCun *et al.*, 2015]. Le principe de l'apprentissage profond est de modéliser des données avec un haut niveau d'abstraction à partir d'une architecture composée de transformations non-linéaires. L'utilisation des réseaux de neurones est alors la méthode utilisée en cela qu'elle est basée sur la succession de plusieurs couches de neurones dans lesquelles, couche après couche, le niveau d'abstraction de représentation des données augmente. Chaque neurone communique alors avec l'ensemble des neurones de la couche précédente dont les valeurs en sortie sont pondérées puis ensuite additionnées dans le neurone et passées au travers une fonction d'activation qui génère la pondération pour la couche suivante. Un apprentissage supervisé ou non-supervisé est alors possible selon les données disponibles. Lors de l'apprentissage supervisée, le poids des synapses est alors ajusté à l'élément ciblé. Ces méthodes ont été conceptualisées depuis de nombreuses années (par Lettvin & al. [Lettvin *et al.*, 1959] pour le neurone artificiel et dans les années 80 pour les algorithmes d'analyses discriminantes [Ackley *et al.*, 1985]) mais ces méthodes nécessitent d'importantes ressources de calcul qui ont rendu leur utilisation impossible pendant de nombreuses années. Aujourd'hui, avec l'accroissement des capacités de calculs des ordinateurs, cette contrainte est beaucoup plus faible et il devient possible d'utiliser cette approche pour réaliser de nombreuses tâches de traitement du signal. Développées depuis 2010 pour les contenus audio (musique, voix), ces méthodes sont de plus en plus utilisées pour des sons

environnementaux dans des tâches de détection [Cakir *et al.*, 2015, Adavanne et Virtanen, 2017] et de classification [Piczak, 2015, Salamon et Bello, 2017] avec des performances supérieures aux méthodes utilisées jusqu'à présent basées sur des paramètres d'extractions et de classifications. Leur utilisation pour la séparation de sources pour des sons environnementaux est encore peu utilisée là où des applications pour la voix [Nugraha *et al.*, 2016] et la musique existent déjà [Huang *et al.*, 2014]. Il serait donc nécessaire et intéressant mettre un place un algorithme basé sur un apprentissage profond basé sur un réseau de neurones et de le confronter aux corpus de sons présents pour comparer ses performances à celles de la NMF Initialisée Seuillée.

Extension à d'autres sources sonores

Enfin, si ces travaux se sont intéressés au bruit du trafic routier, il convient de les étendre à d'autres sources comme la voix et les oiseaux, deux autres sources sonores prépondérantes dans la modélisation de la perception des ESU. Des premières investigations ont commencé dont les premiers résultats semblent induire la nécessité d'adapter les représentations des spectrogrammes selon les sources : si la représentation en bandes de tiers d'octaves est adaptée pour le trafic routier, celle-ci semble moins l'être pour la voix et pour les oiseaux. Un représentation en bandes mel ou en bandes fines mais centrées autour des sources sonores sont des pistes à explorer.

L'étude entreprise durant cette thèse sur l'estimation de la contribution sonore du trafic routier en ville engage donc la réalisation d'autres travaux pouvant être menés à court et moyen termes et ouvrent ainsi des perspectives vers une connaissance plus précise des environnements sonores urbains et des sources qui les composent.

Appendices

Annexe A

Développement de la contrainte de régularité temporelle pour $\beta = 0$ et $\beta = 2$

Inspiré par l'article [Essid et Févotte, 2013] et [Févotte et Idier, 2011], les algorithmes de mise à jour de la NMF avec une contrainte de régularité temporelle sont développés. Initialement développée pour la divergence K-L, ils sont étendus au cas de la distance EUC et de la divergence I-S. Les détails des calculs menant aux algorithmes sont présentés ici. Ils n'ont pas été implémentés et utilisés dans le cadre des travaux de la thèse.

A.1 Cas de la distance Euclidienne

On pose le calcul de la fonction de coût pour la NMF avec une distance EUC et la contrainte :

$$C(\mathbf{H}) = D_2(\mathbf{V}\|\mathbf{WH}) + \alpha_t C_{MM}(\mathbf{H}) \tag{A.1}$$

$$= \frac{1}{2}\sum_n (\mathbf{v}_n - \mathbf{W}\mathbf{h}_n)^2 + \alpha\lambda L(\mathbf{h}_n; \mathbf{h}_{(n+1)}, \mathbf{h}_{(n-1)}) \tag{A.2}$$

avec $\lambda = \|w\|$, la terme de normalisation des matrices $\mathbf{W}$ et α la pondération de la contrainte. On définit $\mathbf{h}$, la valeur de $\mathbf{h}$ à déterminer à l'itération $i+1$, $\tilde{\mathbf{h}}$, sa valeur à l'itération i et $\tilde{\mathbf{v}}_f = \left[\mathbf{W}\tilde{\mathbf{h}}\right]$. En suivant la méthode de *majorisation-minimisation*, une fonction auxiliaire $G_{SM}(\mathbf{h}_k|\tilde{\mathbf{h}}_k)$ est définie :

$$G_{SM}(\mathbf{h}_n|\tilde{\mathbf{h}}_n) = G_2(\mathbf{h}_n|\tilde{\mathbf{h}}_n) + \alpha_t L(\mathbf{h}_n; \mathbf{h}_{(n+1)}, \mathbf{h}_{(n-1)}) \tag{A.3}$$

$$G_2(\mathbf{h}_n|\tilde{\mathbf{h}}_n) = \frac{1}{2}\sum_f\sum_k \frac{w_{fk}\tilde{h}_{kn}v_{fn}^2}{\tilde{v}_{fn}} - 2w_{fk}h_{kn}v_{fn} + w_{fk}\frac{h_k^2}{\tilde{h}_k}\tilde{v}_{fn} \tag{A.4}$$

$$L(\mathbf{h}_n;\mathbf{h}_{(n+1)},\mathbf{h}_{(n-1)}) = \frac{1}{2}\sum_k \lambda_k^2\left[\left(h_{k(n+1)} - h_{kn}\right)^2 + \left(h_{kn} - h_{k(n-1)}\right)^2\right] \tag{A.5}$$

$$= \sum_k \lambda_k^2\left[h_{kn}^2 - h_{kn}\left(h_{k(n+1)} + h_{k(n-1)}\right) + \frac{1}{2}\left(h_{k(n+1)}^2 + h_{k(n-1)}^2\right)\right]. \tag{A.6}$$

La minimisation de la fonction $G_{SM}(\mathbf{h}_n|\tilde{\mathbf{h}}_n)$ est alors déterminée en trouvant les zéros de son gradient $\nabla_{h_k}G_{SM}(\mathbf{h}_n|\tilde{\mathbf{h}}_n)$:

$$\nabla_{h_k}G_{SM}(\mathbf{h}_n|\tilde{\mathbf{h}}_n) = \nabla_{h_k}G_2(\mathbf{h}_n|\tilde{\mathbf{h}}_n) + \alpha_t\nabla_{h_k}L(\mathbf{h}_n;\mathbf{h}_{(n+1)},\mathbf{h}_{(n-1)}) \tag{A.7}$$

$$= \sum_f\sum_k \left(2\alpha_t\lambda_k^2 + \frac{w_{fk}v_{fn}}{\tilde{h}_{kn}}\right)h_{kn} - \left(w_{fk}v_{fn} + \alpha_t\lambda_k^2\left(h_{k(n+1)} + h_{k(n-1)}\right)\right) \tag{A.8}$$

$$= \frac{1}{2}\sum_f\sum_k -2w_{fk}v_{fn} + 2w_{fk}v_{fn}\frac{h_{kn}}{\tilde{h}_{kn}} + \alpha_t\lambda_k^2\left(2h_{kn} - (h_{k(n+1)} + h_{k(n-1)})\right) \tag{A.9}$$

L'équation A.9 est alors résolue et permet d'obtenir l'expression des mises à jour de h_{kn} :

$$\boxed{h_{kn} = \frac{w_{fk}v_{fn} + \alpha_t\lambda_k^2\left(\tilde{h}_{k(n+1)} + \tilde{h}_{k(n-1)}\right)}{2\alpha_t\lambda_k^2 + \frac{w_{fk}v_{fn}}{\tilde{h}_{kn}}}} \tag{A.10}$$

pour $n \in \{2, N-1\}$. Dans le cas où $\alpha_t = 0$, on retrouve bien l' algorithme 3.32. Pour $n = 1$, on obtient :

$$h_{k1} = \frac{w_{fk}v_{f1} + \alpha_t\lambda_k^2\tilde{h}_{k2})}{2\alpha_t\lambda_k^2 + \frac{w_{fk}v_{f1}}{\tilde{h}_{k1}}}. \tag{A.11}$$

Et pour $n = N$, l'équation A.10 devient

$$h_{kN} = \frac{w_{fk}v_{fN} + \alpha_t\lambda_k^2\tilde{h}_{kN}}{2\alpha_t\lambda_k^2 + \frac{w_{fk}v_{fN}}{\tilde{h}_{kN}}}. \tag{A.12}$$

Pour illustrer l'impact de la pondération, une scène de 30 secondes est générée comprenant un passage de voiture dans l'intervalle temporelle [0, 16] secondes suivi de sifflements d'oiseaux dans l'intervalle temporelle [16, 30] secondes. Les deux classes de sons ne présentent pas de recouvrement. Le dictionnaire $\mathbf{W}$ est composé de 5 éléments *trafic* et 5 éléments *oiseaux* construit

à partir de 5 échantillons pour chaque classe de son et avec $w_t = all$. La NMF supervisée ($\beta = 2$, $w_t = all$ et $K = 10$) est appliquée sur cette scène. On illustre en Figure A.1 l'impact de cette contrainte pour $\alpha_t = \{0,\ 2\}$ sur l'évolution du niveau sonores. Le signal trafic estimé doit, en théorie, ne pas être activé lorsque les oiseaux chantent. Or, on observe que dans le cas où la contrainte n'est pas présente, que cette activation est forte. La présence de la contrainte permet de réduire cette activation. Durant le signal *trafic*, l'effet de la contrainte s'observe en lissant l'évolution du signal *trafic*. La régularité temporelle permet d'estimer plus une enveloppe temporelle qu'une estimation très précise (comme dans le cas où $\alpha_t = 0$).

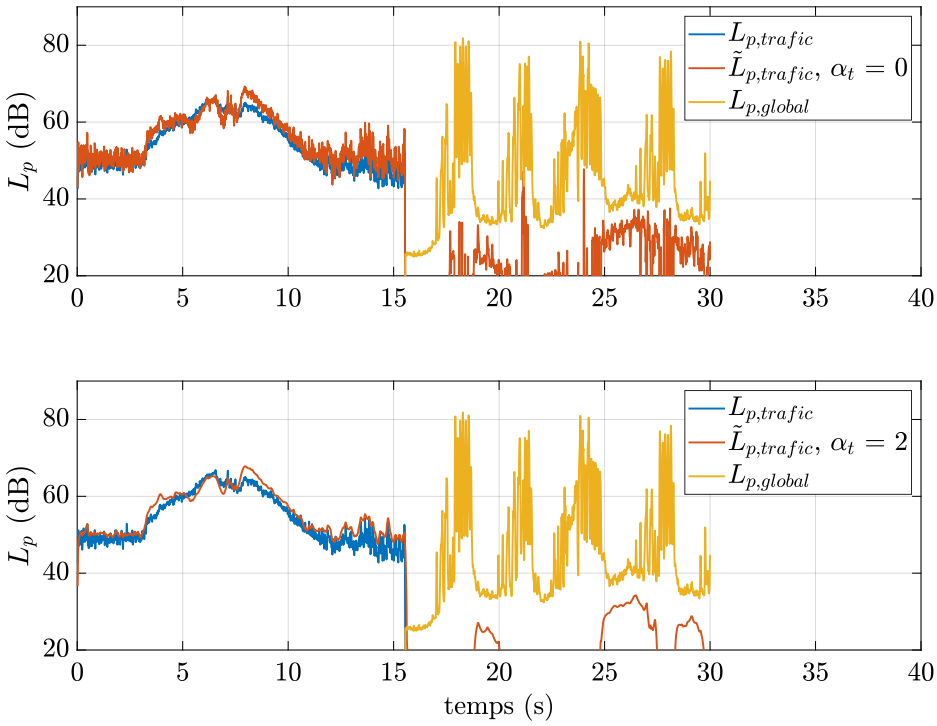


FIGURE A.1 – Influence de la contrainte de régularité temporelle pour $\beta = 2$ selon l'algorithme obtenu par *majorisation-minimisation*.

A.2 Cas de la divergence d'Itakura-Saito

Dans le cas d'une divergence I-S, le problème est similaire avec

$$C(\mathbf{H}) = D_0(\mathbf{V}\|\mathbf{WH}) + \alpha_t C_{MM}(\mathbf{H}) \tag{A.13}$$

$$= \sum_k \frac{\mathbf{v}_n}{\mathbf{W}\mathbf{h}_n} - \log \frac{\mathbf{v}_n}{\mathbf{W}\mathbf{h}_n} - 1 + \alpha_t L(\mathbf{h}_n; \mathbf{h}_{(n+1)}, \mathbf{h}_{(n-1)}) \tag{A.14}$$

La fonction auxiliaire $G_0(\mathbf{h}_n|\tilde{\mathbf{h}}_n)$ n'est toutefois pas la même :

$$G_0(\mathbf{h}_n|\tilde{\mathbf{h}}_n) = \sum_f \sum_k \left[\frac{w_{fk}\tilde{h}_{kn}}{\tilde{v}_{fn}} v_{fn} \frac{\tilde{h}_{kn}}{\tilde{v}_{fn} h_{kn}} + \log \tilde{v}_{fn} + \frac{1}{\tilde{v}_{fn}} (w_{fk}(h_{kn} - \tilde{h}_{kn}) + v_f(\log v_{fn} - 1) \right] \tag{A.15}$$

$$\nabla_{h_k} G_{SM}(\mathbf{h}_n|\tilde{\mathbf{h}}_n) = \nabla_{h_k} G_0(\mathbf{h}_n|\tilde{\mathbf{h}}_n) + \alpha_t \nabla_{h_k} L(\mathbf{h}_n; \mathbf{h}_{(n+1)}, \mathbf{h}_{(n-1)}) \tag{A.16}$$

$$= \sum_f \sum_k -w_{fk} \frac{\tilde{h}_{kn}^2}{h_{kn}^2} \frac{v_{fn}}{\tilde{v}_{fn}^2} + \frac{w_{fk}}{\tilde{v}_{fn}} + \alpha_t \lambda_k^2 \left[2h_{kn} - \left(h_{k(n+1)} + h_{k(n-1)} \right) \right] \tag{A.17}$$

$$= \alpha_t \lambda^2 \sum_f \sum_k 2h_{kn}^3 + \left[\frac{w_{fk}}{\tilde{v}_{fn}} - \alpha_t \lambda_k^2 \left(h_{k(n+1)} + h_{k(n-1)} \right) \right] h_{kn}^2 - w_{fk} \frac{v_{fn}}{\tilde{v}_{fn}^2} \tilde{h}_{kn}^2 \tag{A.18}$$

On obtient un problème de type équation du 3e ordre : $ax^3 + bx^2 + cx + d = 0$ avec

— $a = 2\alpha_t \lambda^2$,

— $b = \frac{w_{fk}}{\tilde{v}_f} - \alpha_t \lambda_k^2 \left(h_{k(n+1)} + h_{k(n-1)} \right)$,

— $c = 0$,

— $d = -w_{fk} \frac{v_{fn}}{\tilde{v}_{fn}} \tilde{h}_{kn}^2$.

Cette polynôme d'ordre 3 se résout à l'aide de la méthode de Cardan [Nickalls, 1993], qui dans le cas présent, s'exprime sous la forme :

$$x^3 + \frac{b}{a}x^2 + \frac{d}{a} = 0. \tag{A.19}$$

Dans un premier temps, un premier changement de variable est effectué $x = X - \frac{b}{3a}$ et on obtient :

$$X^3 + pX + q = 0 \tag{A.20}$$

avec $p = -\frac{b^2}{3a^2}$ et $q = \frac{2b^3}{27a^3} + \frac{d}{a}$. La variable X est alors décomposée en deux variables complexes : $X = u + v$. L'équation A.20 devient alors :

$$u^3 + v^3 + (3uv + p)(u + v) + q = 0. \tag{A.21}$$

L'équation A.21, pour être résolue, implique alors deux conditions :

$$3uv = -p, \tag{A.22a}$$

$$u^3 + v^3 = -q. \tag{A.22b}$$

Du système d'équations A.22, on obtient :

$$u^3 + v^3 = -q, \tag{A.23a}$$

$$u^3 v^3 = -\frac{p^3}{27}. \tag{A.23b}$$

Si on réalise un nouveau changement de variable $u^3 = U$ et $v^3 = V$, on exprime le système d'équation A.23 comme des polynômes de second degré :

$$U^2 + qU - \frac{p^3}{27} = 0, \tag{A.24a}$$

$$V^2 + qV - \frac{p^3}{27} = 0. \tag{A.24b}$$

Le système A.24 se résout classiquement :

$$\Delta_U = \Delta_V = \Delta = q^2 - 4\frac{p^3}{27}. \tag{A.25}$$

Les équations A.24 ont alors les deux mêmes solutions. Les valeurs de X et solutions de l'équation A.20 dépendent alors du signe de Δ :

— pour $\Delta > 0$,

$$X_1 = \sqrt[3]{\frac{-q + \sqrt{\Delta}}{2}} + \sqrt[3]{\frac{-q - \sqrt{\Delta}}{2}}, \tag{A.26a}$$

$$X_2 = j\sqrt[3]{\frac{-q + \sqrt{\Delta}}{2}} + j^2\sqrt[3]{\frac{-q - \sqrt{\Delta}}{2}}, \tag{A.26b}$$

$$X_3 = j^2\sqrt[3]{\frac{-q + \sqrt{\Delta}}{2}} + j\sqrt[3]{\frac{-q - \sqrt{\Delta}}{2}} \tag{A.26c}$$

avec $j = e^{2i\pi/3}$. Une autre forme d'écriture la solution est possible sous une forme trigonométrique :

$$X_z = 2\sqrt{\frac{-p}{3}}\cos\left(\frac{1}{3}\arccos\left(\frac{-q}{2}\sqrt{\frac{27}{-p^3}}\right) + \frac{2(z-1)\pi}{3}\right) \tag{A.27}$$

avec $z \in \{1, 2, 3\}$.

— Pour $\Delta = 0$, parmi les 3 solutions possibles, 1 seule solution est réelle :

$$X_1 = \frac{3q}{p}, \tag{A.28a}$$

$$X_2 = X_3 = \frac{-3q}{2p}. \tag{A.28b}$$

— Pour $\Delta < 0$, les équations A.24 ont pour solutions

$$S_1 = \frac{-q + i\sqrt{|\Delta|}}{2}, \tag{A.29}$$

$$S_2 = \frac{-q - i\sqrt{|\Delta|}}{2}. \tag{A.30}$$

L'équation A.20 possède alors une solution réelle et 2 solutions complexes qui sont :

$$X_1 = \sqrt[3]{S_1} + \sqrt[3]{S_2}, \tag{A.31a}$$

$$X_2 = j\sqrt[3]{S_1} + j^2\sqrt[3]{S_2}, \tag{A.31b}$$

$$X_3 = j^2\sqrt[3]{S_1} + j\sqrt[3]{S_2}. \tag{A.31c}$$

Ainsi, la solution du problème A.18 est donc la valeur X_z qui est réelle et positive parmi les trois solutions possibles et définie selon le signe de Δ :

$$\boxed{h_{kn} = X_z^+ - \tfrac{b}{3a}.} \tag{A.32}$$

Le même exemple illustratif que pour $\beta = 2$ est repris ici en Figure A.2. On relève que la valeur α_t est bien plus élevée pour la divergence IS ($\alpha_t = 10^7$) que pour la distance EUC en raison de la valeur d et du rapport $\frac{w_{fk}}{\bar{v}_f}$ de b, dans l'équation du 3$^{\text{e}}$ ordre, qui prennent des valeurs élevés.

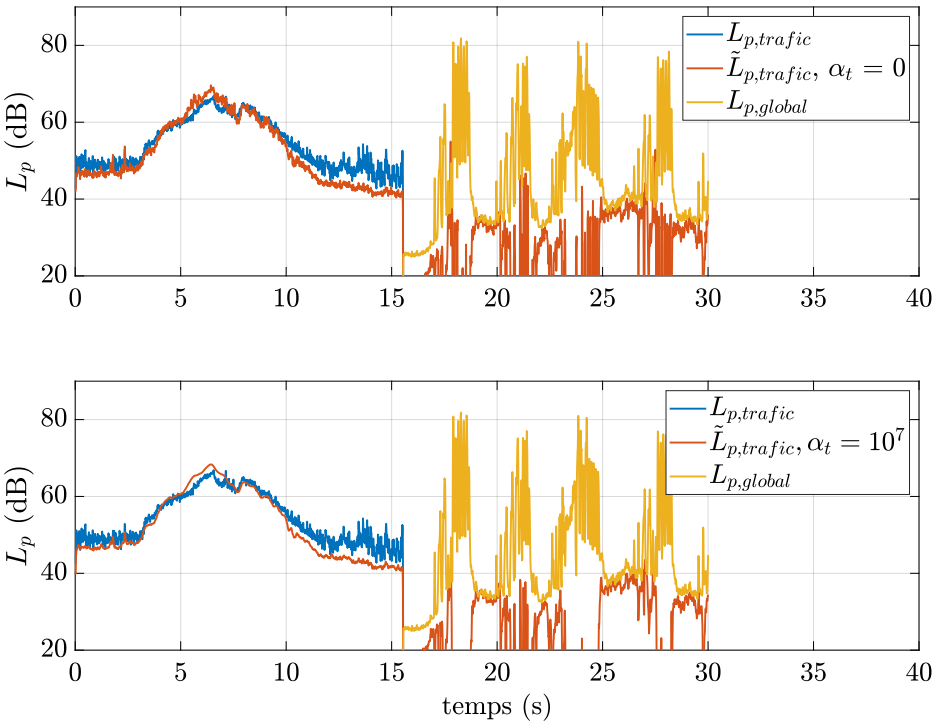


FIGURE A.2 – Influence de la contrainte de régularité temporelle pour $\beta = 0$ selon l'algorithme obtenu par *majorisation-minimisation*.

Annexe B

Analyses complémentaires des résultats du test perceptif

B.1 Effets des auditeurs sur l'évaluation des scènes

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

C'est ainsi une ANOVA à deux facteurs (*type de scènes* (*enregistré*, *répliqué*) et *auditeur* (50 auditeurs)) avec interaction qui est considérée. De la même manière que le test de Student, l'ANOVA est un outil statistique qui permet de comparer des moyennes d'échantillons et d'étudier l'effet des variables qualitatives (ou facteurs), pouvant prendre plusieurs valeurs (ou niveaux), sur une variable quantitative. Dans ce test, deux hypothèses sont émises sur les distributions :

— les distributions des échantillons des différentes catégories sont semblables (hypothèse *nulle* H_0),

— les distributions sont différentes, (hypothèse *alternative* H_1).

Pour cela, la statistique de test est la statistique de Fischer F. De ces indices, la valeur p établit, là encore, la probabilité d'obtenir sous l'hypothèse nulle, des résultats aussi extrêmes que ceux observés. Les résultats sont résumés dans le Tableau B.1. On y résume la somme des carrés des écarts (SCE), les degrés de liberté du système (DDL), la variance, la statistique F et la valeur p.

Le facteur *auditeur* a une influence significative (valeur $p < \alpha$) révélant que les auditeurs n'ont pas les mêmes échelles d'évaluation. L'influence du facteur *type* reste toujours non significative. Son interaction avec le facteur *auditeur* est non-significatif également. Le phénomène

Tableau B.1 – Résultats de l'ANOVA avec interaction avec les facteurs *type* et *auditeur*.

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

d'interaction traduit l'influence des différents niveaux d'un facteur sur l'autre facteur. Ici, l'interaction entre le facteur *auditeur* et *type* est non-significative, ce qui signifie que pour chaque juge, la perception des scènes est similaire, même si entre chaque juge des dissimilarités existent.

B.2 Effets de l'ambiance sonore

Une seconde ANOVA à deux facteurs avec interaction est effectuée avec pour facteur le *type* (*enregistrée*, *répliquée*) et l'*ambiance sonore* (*Parc, Rue calme, Rue bruyante, Rue très bruyante*) afin de déterminer si la perception du réalisme est différente selon l'ambiance sonore. Les résultats de l'ANOVA sont résumés dans le Tableau B.2.

Tableau B.2 – Résultats de l'ANOVA avec interaction avec les facteurs *type* et *ambiance*.

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

Si l'impact du facteur *type* est toujours non significatif, celui du facteur *ambiance* et l'interaction entre les deux facteurs sont toutefois significatifs ($p < \alpha$). L'influence principale du facteur *ambiance* signifie qu'il y a une distinction entre les distributions des notes selon l'ambiance sonore. Le phénomène d'interaction traduit le fait que l'effet du type de scène (enregistré - répliqué) sur le réalisme varie en fonction de l'ambiance considérée.

Pour visualiser ce phénomène d'interaction entre le type de scènes et l'ambiance sonore, l'évolution de la note moyenne dans chaque cas est tracée (Figure B.1). On observe que selon l'ambiance sonore, la note de réalisme des scènes répliquées peut être inférieure ou supérieure par rapport aux scènes enregistrées. Cette évolution traduit une interaction croisée dont l'origine est toutefois difficile à estimer.

Même si les moyennes globales et les distributions entre les scènes enregistrées et répliquées sont similaires, des disparités existent selon les auditeurs ou les ambiances sonores sans toutefois que celles-ci remettent en cause les similarités entre les deux types.

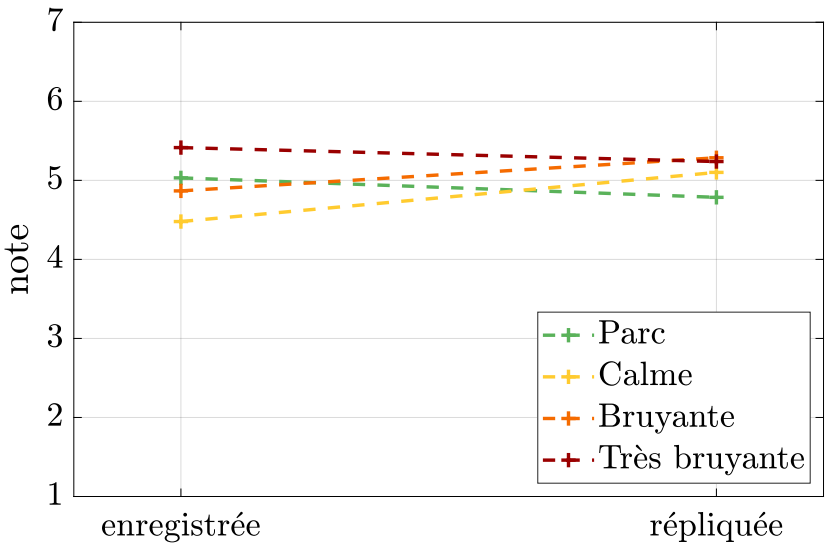


FIGURE B.1 – Diagramme des effets d'intéraction type∗ambiance.

Annexe C

Estimation du seuil optimal par des indicateurs sonores

Dans la partie 6.3, les seuils optimaux $t_{h,opt}$ par ambiance sonore ont été définis en considérant la NMF IS optimale avec $\beta = 2$, $K = 300$ et $w_t = 1$ s. Ces seuils et les erreurs correspondantes sont résumés dans le Tableau C.1 et peuvent être considérés dans les réseaux de capteurs fixes où l'ESU et la position des capteurs peuvent être définis au préalable. Toutefois, leur prise en compte a un impact relativement faible sur les erreurs MAE_g.

Cependant, une étude a été réalisée en vue de, non plus définir un seuil fixe selon l'ambiance sonore, mais de définir un seuil adaptatif basé sur des indicateurs physiques corrélés avec l'évolution des seuils optimaux. Cette adaptation permettrait de définir un seuil en fonction des ESU qui sont susceptibles d'évoluer en fonction de la période de l'année ou de la journée.

Tableau C.1 – Erreurs MAE_{60} minimales selon le seuil fixe t_h et optimal $t_{h,opt}$ par ambiance sonore.

ambiance	**Parc**	**Rue calme**	**Rue bruyante**	**Rue très bruyante**	MAE_{60} (dB)
seuil fixe t_h	0,35	0,35	0,35	0,35	0,35
erreur MAE_{60}	2,13 (± 3,84)	1,62 (± 1,85)	0,57 (± 0,54)	0,32 (± 0,20)	1,16 (± 0,86)
seuil optimal $t_{h,opt.}$	0,38	0,35	0,33	0,31	-
erreur MAE_{60}	2,03 (± 3,47)	1,62 (± 1,85)	0,56 (± 0,67)	0,28 (± 0,31)	1,12 (± 0,83)

Des outils de classification pour les environnements sonores existent déjà dans la littérature et sont basés sur des indicateurs simples [Can et Gauvreau, 2015], [Rychtáriková et Vermeir, 2013] ou sur des outils plus complexes comme dans [Salamon et Bello, 2015]. Dans le cadre de ces travaux, il est nécessaire que ces indicateurs soient simples en vue de les considérer sur l'ensemble des types de mesures possibles (fixes, mobiles, participatives) et dont les puissances de calculs sont limités. Ainsi, nous considérons une approche simple basée sur l'estimation d'indicateurs de niveaux sonores. Plusieurs indicateurs sont considérés comme des niveaux sonores par bandes de tiers d'octave (L_{500}, L_{1k}, L_{2k}, L_{5k}) ainsi que des niveaux sonores fractiles (L_{10}, L_{50}, L_{90}). Ces derniers expriment les niveaux sonores dépassés pendant un pourcentage de temps défini. Ainsi

L_{10} exprime le niveau sonore dépassé 10 % du temps, cela résume les hauts niveaux sonores, à l'inverse l'indice L_{90} traduit le niveau sonore dépassé 90 % du temps et exprime donc l'équivalent du bruit de fond sonore d'une scène.

Toutefois, cette approche se heurte à la calibration des scènes sonores. En effet, les enregistrements qui ont servi à la construction du corpus *SOUR* n'ont pas été accompagnés d'un fichier de calibration. Il n'a donc pas été possible d'estimer leur niveau sonore exact et donc de calibrer les scènes sonores simulées du corpus. Leurs niveaux sonores absolus ne peuvent donc pas être estimés. Une approximation avait été réalisée dans le Tableau 4.8 à partir des informations renseignées dans la Figure 4.7, dans le but d'harmoniser les fichiers audio par ambiances, mais celle-ci n'est pas suffisante ici. En revanche, si les niveaux absolus de ces indicateurs ne peuvent pas être estimés par défaut de calibration, en considérant que les scènes du corpus *SOUR* sont représentatives des enregistrements sonores urbains, le rapport de ces niveaux sonores entre eux reste correct. Ainsi les différences entre les niveaux par bandes de tiers d'octave et entre les niveaux fractiles, exprimées en dB, peuvent être considérées :

$$\Delta_{L_{500}-L_{1k}} = L_{500} - L_{1k}, \tag{C.1a}$$
$$\Delta_{L_{500}-L_{2k}} = L_{500} - L_{2k}, \tag{C.1b}$$
$$\Delta_{L_{500}-L_{5k}} = L_{500} - L_{5k}, \tag{C.1c}$$
$$\Delta_{L_{1k}-L_{2k}} = L_{1000} - L_{2k}, \tag{C.1d}$$
$$\Delta_{L_{1k}-L_{5k}} = L_{1000} - L_{5k}, \tag{C.1e}$$
$$\Delta_{L_{2k}-L_{5k}} = L_{500} - L_{5k}, \tag{C.1f}$$
$$\Delta_{L_{10}-L_{50}} = L_{10} - L_{50}, \tag{C.1g}$$
$$\Delta_{L_{10}-L_{90}} = L_{10} - L_{90}, \tag{C.1h}$$
$$\Delta_{L_{50}-L_{90}} = L_{50} - L_{90}. \tag{C.1i}$$

La Figure C.1 résume les évolutions du seuil optimal et des indicateurs $\Delta_{L_x-L_y}$ par ambiance sonore. À partir de ces valeurs, les corrélations entre l'évolution du seuil optimal par ambiance sonore et celles des indicateurs sont alors calculées pour déterminer celui qui évolue de la même manière que $t_{h,opt}$. Les valeurs absolues des corrélations sont résumées dans le Tableau C.2.

On obtient une forte corrélation pour les indicateurs $\Delta_{L_{1k}-L_{5k}}$, $\Delta_{L_{500}-L_{5k}}$ et $\Delta_{L_{2k}-L_{5k}}$. Avec une différence entre un niveau situé dans les basses (500 Hz) ou moyennes fréquences (1 kHz et 2 kHz) avec un niveau situé dans les plus hautes fréquences (5 kHz), on réussit à définir les environnements sonores. En effet, les ambiances *Parc* et *Rue calme* avec la présence d'oiseaux sont susceptibles de contenir plus de hautes fréquences alors que la source *trafic* est moins présente. À l'inverse, *Rue bruyante* et *Rue très bruyante* possèdent une plus grande proportion de basses fréquences. Les indicateurs basés sur la différence des niveaux fractiles et celui basé sur la différence entre les niveaux sonores des bandes de 500 Hz et de 2 kHz sont ceux obtenant les plus faibles corrélations. En se référant à la Figure C.1, là où les autres indicateurs suivent une évolution quasi linéaire, similaire à celle des seuils optimaux, ces 4 indicateurs ont des valeurs

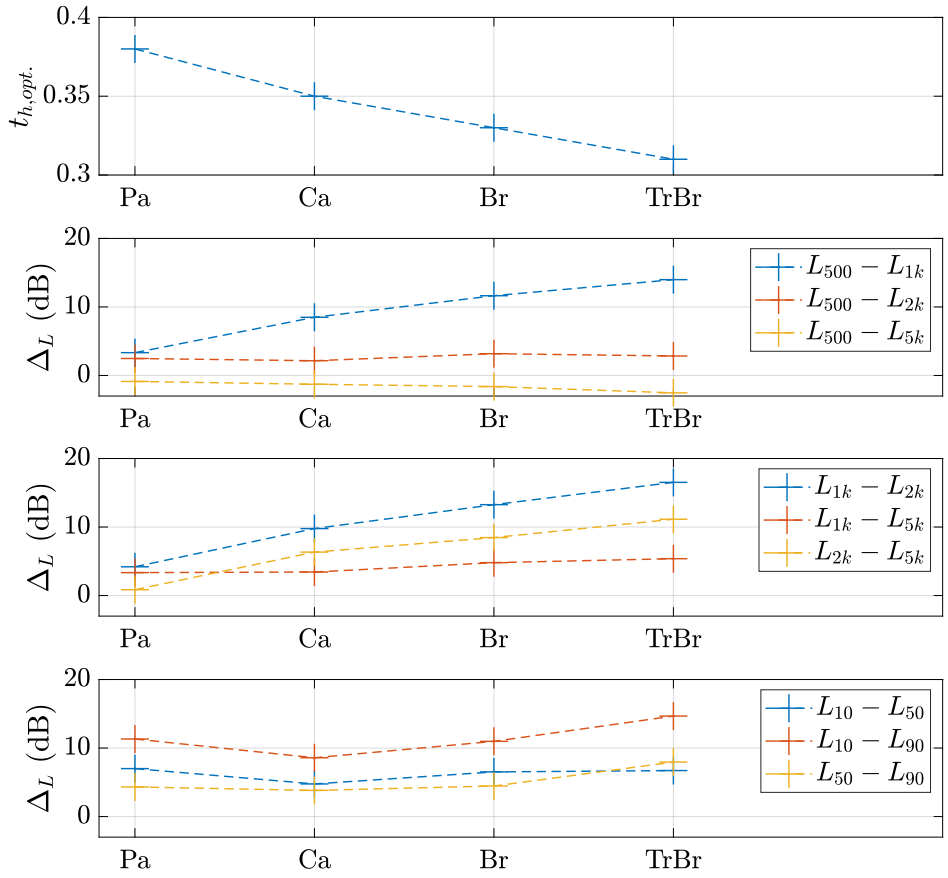


FIGURE C.1 – Évolution des seuils optimaux et des indicateurs Δ_L selon les ambiances.

plus faibles pour l'ambiance *Rue calme*, ce qui diminue leur corrélation.

À partir des valeurs seuils $t_{h,opt}$, et des valeurs des indicateurs par ambiance sonore, des régressions linéaires sont estimées afin de relier la valeur d'un indicateur $\Delta_{L_x - L_y}$ calculé sur une scène à un seuil interpolé $t_{h,interp.}$. Le principe de la méthode est présenté par un schéma en Figure C.3. La NMF IS ($\beta = 2$, $K = 300$, $w_t = 1$ s) est de nouveau appliquée sur le corpus *SOUR* où la valeur du seuil est trouvée par interpolation (voir Figure C.4). Toutefois, si l'indicateur $\Delta_{L_x - L_y}$ de la scène excède les valeurs limites déterminées aux ambiances *Parc* et *Rue très bruyante*, les seuils prendront les valeurs optimales $t_{h,opt.}$ de ces ambiances afin de ne pas réaliser d'extrapolation.

Les erreurs MAE_g obtenues sur le corpus sont résumées dans le Tableau C.2. Ce sont les approches basées sur les indicateurs $\Delta_{L_{1k} - L_{5k}}$ et $\Delta_{L_{2k} - L_{5k}}$ qui génèrent les plus faibles erreurs. Celles-ci sont également plus faibles que celles basées sur le seuil fixe et sur les seuils optimisés. Ces deux indicateurs sont ceux ayant les plus fortes corrélations avec l'évolution des seuils $t_{h,opt}$. La fonction d'interpolation pour $\Delta_{L_{1k} - L_{5k}}$ est alors, sur l'intervalle $\Delta_{L_{1k} - L_{5k}} \in [1,22;\ 15,84]$:

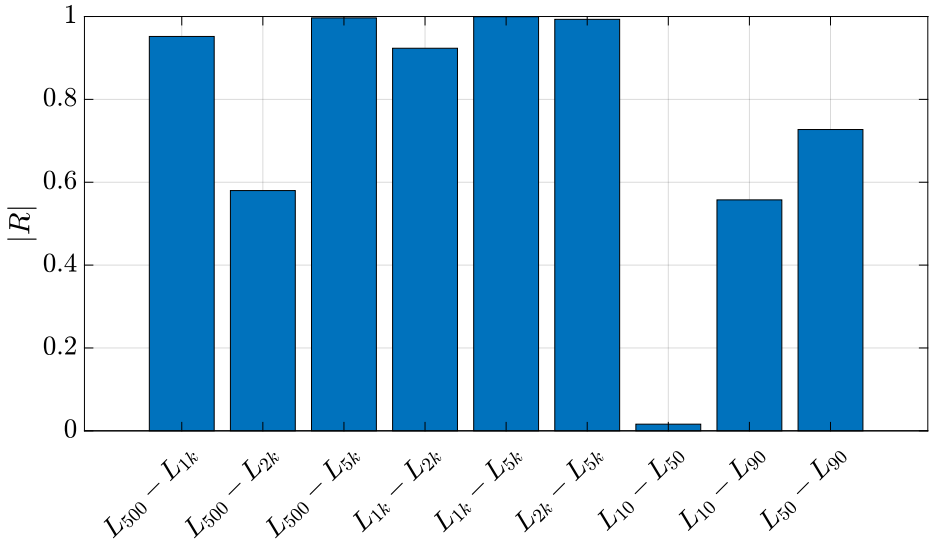


FIGURE C.2 – Corrélations entre l'évolution des seuils optimaux et des indicateurs $\Delta_{L_x - L_y}$.

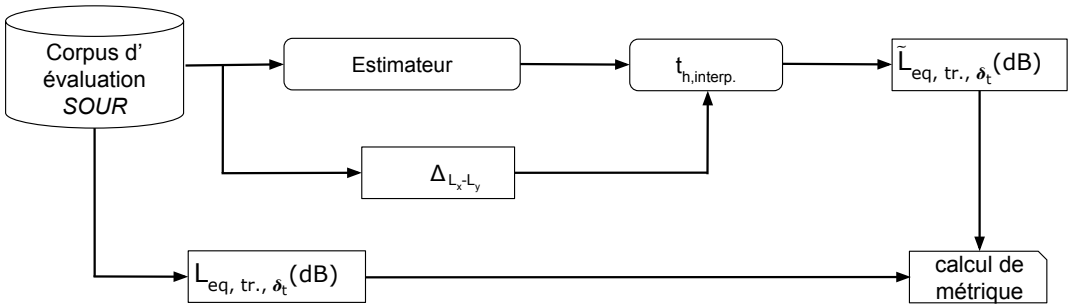


FIGURE C.3 – Schéma de principe de la détermination du niveau sonore du trafic par seuil interpolé.

$$t_{h,interp.} = -5,70 \times 10^{-3} \times \Delta_{L_{1k} - L_{5k}} + 0,40. \tag{C.2}$$

La Figure C.5 représente les erreurs par ambiance sonore pour la NMF IS basées sur le seuil fixe $t_h = 0,35$, les seuils optimisés $t_{h,opt.}$ et sur les seuils interpolés $t_{h,interp.}$. S'il y a bien une diminution de l'erreur, celle-ci reste relativement limitée. L'ambiance *Parc* est le seul cas où le seuil déduit de l'interpolation génère des erreurs supérieures aux deux autres approches. Cette particularité peut être due à la sensibilité de l'erreur plus forte dans cette ambiance (voir partie 6.6.2 et la Figure 6.15). Dans les autres ambiances, la déduction du seuil $t_{h,interp.}$ par interpolation réduit les erreurs.

Cette approche est une première piste d'ouverture pour améliorer l'estimation du niveau sonore du trafic, elle reste à être validée sur des corpus plus grands et plus variés afin d'être ajustée. Mais cette proposition reste intéressante en vue d'adapter la méthode aux divers en-

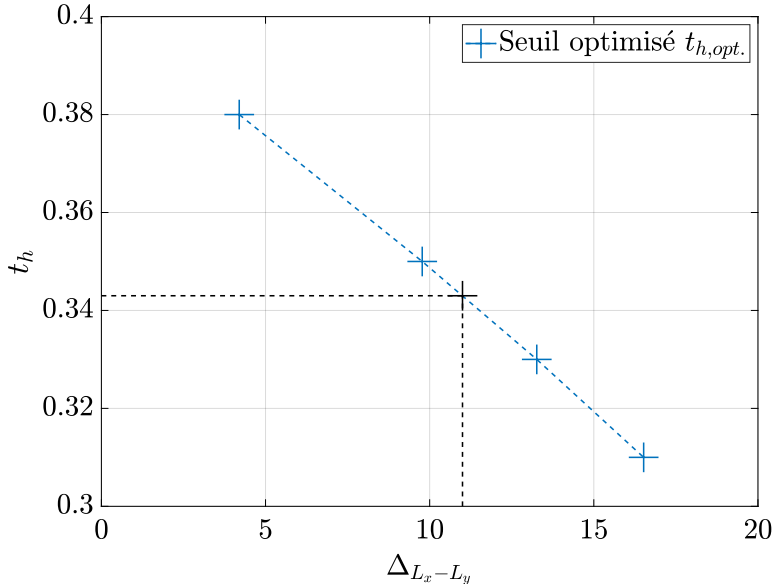


FIGURE C.4 – Exemple d'interpolation réalisée pour une scène avec un indicateur $\Delta_{L_{1k}-L_{5k}} = 11$: $t_{h,interp} = 0{,}343$.

vironnements sonores : par un seuil variable déduit par des indicateurs sonores de la scène, il permet de mieux prendre en compte les évolutions sur la journée ou sur l'heure des ESU auprès du capteur et donc de limiter les erreurs de la NMF IS sur l'estimation du niveau sonore du trafic.

TABLEAU C.2 – Influence de l'indicateur d'optimisation dans l'estimation de l'erreur MAE_g.

	MAE_g (dB)
seuil fixe t_h	1,16 (± 0,86)
seuil optimisé $t_{h,opt}$	1,12 (± 0,83)
$\mathbf{\Delta_{L_{1k}-L_{5k}}}$	**1,08 (± 0,87)**
$\mathbf{\Delta_{L_{2k}-L_{5k}}}$	**1,09 (± 0,86)**
$\Delta_{L_{500}-L_{5k}}$	1,12 (± 0,90)
$\Delta_{L_{1k}-L_{2k}}$	1,12 (± 0,93)
$\Delta_{L_{500}-L_{1k}}$	1,12 (± 0,95)
$\Delta_{L_{50}-L_{90}}$	1,18 (± 0,92)
$\Delta_{L_{10}-L_{90}}$	1,29 (± 1,09)
$\Delta_{L_{10}-L_{50}}$	1,36 (± 1,05)
$\Delta_{L_{500}-L_{2k}}$	1,37 (± 1,16)

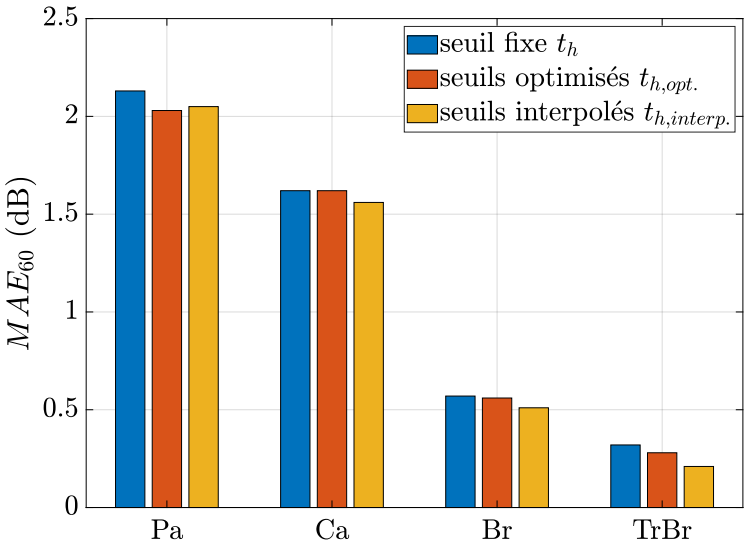


FIGURE C.5 – Influence de la méthode de seuillage (seul fixe t_h, optimisé $t_{h,opt.}$, interpolé $t_{h,interp.}$) à partir de l'indicateur $\Delta_{L_{1k}-L_{5k}}$ sur les erreurs MAE_{60}.

Annexe D

Impact de la contrainte de parcimonie sur la corpus *SOUR*

En plus de la contrainte de régularité temporelle étudiée dans le chapitre 6.6.1, la contrainte de parcimonie a été considérée en fin de thèse. Les résultats obtenus mériteraient donc d'être approfondis, mais les premières valeurs obtenues sont tout de même présentées dans cette annexe. Appliquée sur la matrice $\mathbf{H}$, la contrainte de parcimonie (ou *sparsness*) a pour but de réduire le nombre d'éléments activés à chaque trame temporelle en pénalisant les termes non nuls. De la même façon, cette contrainte se traduit par l'ajout d'un second terme dans l'expression de la fonction de coût :

$$\min D\left(\mathbf{V}\|\mathbf{WH}\right) + \alpha_{sp} C_{sp}(\mathbf{H}) \quad \text{avec} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{D.1}$$

avec $C_{sp}(\mathbf{H})$, la contrainte exprimée dans l'équation 3.52 et α_{sp} la pondération. De même que pour la contrainte de *smoothness*, la parcimonie est appliquée sur les matrices $\mathbf{H}$ de la NMF SUP et IS. Pour la NMF SEM, cette contrainte est appliquée non pas sur $\mathbf{H_s}$ mais sur $\mathbf{H_r}$ afin de mieux inclure des évènements ponctuels dans $\mathbf{W_r}$ et moins les composantes *trafic*. La parcimonie est testée pour plusieurs valeurs de pondération : $\alpha_{sp} \in \{0,05,\ 0,1,\ 0,2,\ 0,5,\ 1\}$. 400 itérations sont également réalisées sur l'ensemble des combinaisons testées. Le Tableau D.1 résume les erreurs MAE_g.

Les résultats pour $\beta = 2$ sur les différentes versions de la NMF ne sont pas présents en raison du comportement de la mise à jour des matrices pour ce cas. En effet, l'algorithme de mise à jour présente au numérateur une soustraction qui peut générer des valeurs négatives selon les valeurs de α_{sp}. Ce phénomène est apparu durant les calculs et n'a pas pu être traité. Toutefois les résultats obtenus pour $\beta \in \{0,\ 1\}$ sont correctes. La contrainte de parcimonie a un impact différent selon la version de la NMF et les valeurs de β. Dans le cas de la NMF SUP, l'ajout de la contrainte de parcimonie n'a qu'un faible impact sur les erreurs MAE_g pour la divergence KL. Dans le cas de la NMF IS, cette contrainte ne permet pas, là encore, d'améliorer les résultats. Cette approche n'est donc pas adaptée à l'ajout de contraintes et est plus performante sans. Une fois encore, l'influence de la contrainte est surtout notable pour la NMF SEM. Si cet impact

TABLEAU D.1 – Erreurs MAE_{60} les plus faibles pour les combinaisons optimales des modalités des estimateurs pour le corpus d'évaluation *SOUR* en présence d'une pondération de parcimonie.

méthode	f_c (kHz)	β	w_t	K	t_h	α_t	MAE_{60} (dB)
filtre PB	20	-	-	-	-	-	3,62 (± 3,93)
	0,5	-	-	-	-	-	1,99 (± 1,37)
NMF SUP	-	2	0,5	25	-	0	2,13 (± 2,22)
	-	0	0,5	200	-	0,5	3,73 (± 4,01)
	-	1	0,5	50	-	0,05	2,50 (± 1,90)
	-	2	-	-	-	-	-
NMF SEM	-	1	0,5	300	-	0	1,93 (± 0,42)
	-	0	2	300	-	0,50	1,99 (± 0,70)
	-	1	0,5	100	-	0,05	1,50 (± 0,89)
	-	2	-	-	-	-	-
NMF IS	-	2	1	300	0,35	0	1,16 (± 0,86)
	-	0	*all*	50	0,10	0,44	1,62 (± 1,02)
	-	1	1	200	0,05	0,40	1,60 (± 0,72)
	-	2	-	-	-	-	-

est faible pour la divergence IS, pour la divergence KL, le gain est significatif avec une erreur MAE_g de 1,50 dB. On observe dans la Figure D.1 les erreurs MAE_{60} pour la NMF SEM ($\beta = 1$, $K = 100$ et $w_t = 0{,}5$ s) avec et sans pondération. On observe un comportement similaire de la méthode soumise à la parcimonie à celui obtenu avec la contrainte de régularité temporelle : une augmentation des erreurs dans *Parc* pour ensuite, avec l'augmentation de la présence de trafic, diminuer de plus en plus. Cette évolution peut être illustrée, là encore, avec le spectre du signal $\mathbf{W_r H_r}$ en Figure D.2. Ici c'est $\mathbf{H_r}$ qui est contraint. Dans le cas de l'ambiance *Parc*, la partie libre de la NMF SEM n'est donc plus en mesure de se comporter aussi librement que sans pondération. On peut alors supposer que certains éléments *trafic*, non utilisés lorsque la pondération est nulle, seront utilisés sur des sources interférantes en vue de minimiser la distance $D(\mathbf{V}\|\mathbf{WH})$, ce qui dégrade l'estimation du trafic. À l'inverse, en présence de trafic, étant contraint d'avoir une parcimonie dans $\mathbf{H_r}$, il est moins aisé pour la NMF SEM d'y inclure des composantes *trafic*, les éléments de $\mathbf{W_s}$ sont alors mieux mis à contribution que dans le cas où la pondération est nulle, ce qui permet alors d'améliorer l'estimation du signal trafic.

La contrainte de parcimonie a donc un impact significatif pour la NMF SEM avec $\beta = 1$, $w_t = 0{,}5$, $K = 100$ et $\alpha_{sp} = 0{,}05$. En contraignant cette fois-ci la partie libre du dictionnaire, on arrive à limiter l'ajout de composante *trafic* dans la partie mobile du dictionnaire et ainsi améliorer l'estimation du niveau sonore du trafic. Cette contrainte engendre toutefois une augmentation des erreurs dans l'ambiance *Parc*. Son impact sur la NMF IS reste, comme la contrainte de régularité temporelle, négatif et dégrade les estimation du niveau sonore du trafic.

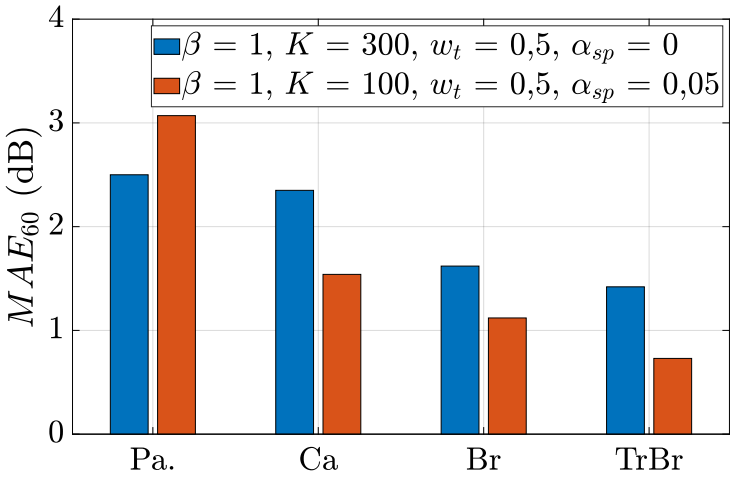


FIGURE D.1 – Influence de la parcimonie sur l'erreur MAE_{60} selon la NMF SEM optimale avec et sans pondération.

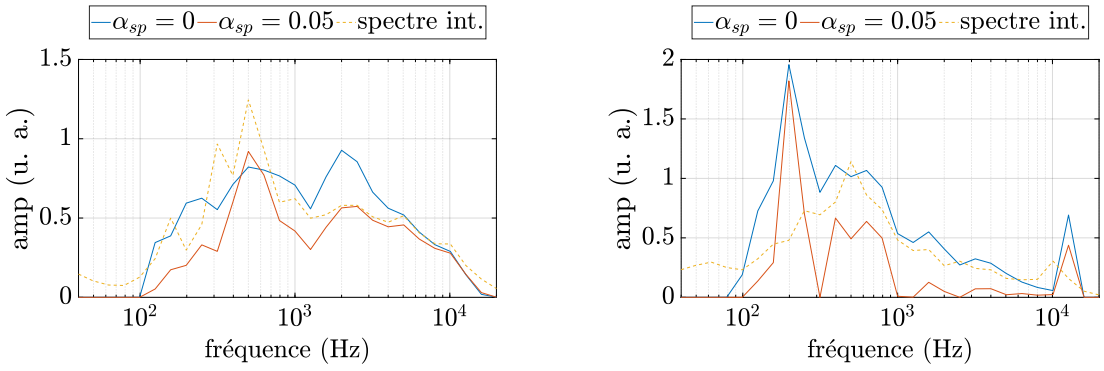


FIGURE D.2 – Comparaison des spectres de la classe *interférante* avec la somme des 2 éléments de $\mathbf{W_r}$ pour le cas sans pondération ($\alpha_{sp} = 0$) et avec ($\alpha_{sp} = 0,05$) pour la scène 2 de l'ambiance *Parc* (a) et la scène 7 de *Rue très bruyante* (b).

Annexe E

Correspondance des noms de scènes du corpus *SOUR*

Dans cette annexe, on résume la correspondance des noms entre les scènes enregistrées du corpus de référence, nommées selon le jour (1 ou 2), la direction du trajet (est-ouest EW ou ouest-est WE) et le numéro sur le parcours (de 01 à 19), et les noms des scènes répliquées sous le logiciel *SimScene*, nommées selon leur ambiance sonore (Parc, Rue calme, Rue bruyante, Rue très bruyante) et un numéro d'identification (01, 02...)

Tableau E.1 – Correspondances des noms des scènes enregistrées et répliquées pour l'ambiance *Parc*.

Ambiance	**n°**	**scène originale**
Parc	01	1-EW-03
	02	1-EW-08
	03	1-EW-10
	04	1-EW-19
	05	2-EW-03
	06	2-EW-19
	07	2-WE-03
	08	2-WE-19

Tableau E.2 – Correspondances des noms des scènes enregistrées et répliquées pour l'ambiance *Rue calme*.

Ambiance	n°	**scène originale**
Rue calme	01	1-EW-04
	02	1-EW-05
	03	1-EW-07
	04	1-EW-11
	05	1-EW-13
	06	1-EW-14
	07	1-EW-17
	08	1-WE-04
	09	1-WE-05
	10	1-WE-06
	11	1-WE-07
	12	1-WE-08
	13	1-WE-11
	14	1-WE-13
	15	1-WE-14
	16	1-WE-17
	17	2-EW-04
	18	2-EW-05
	19	2-EW-06
	20	2-EW-07
	21	2-EW-08
	22	2-EW-11
	23	2-EW-13
	24	2-EW-17
	25	2-WE-04
	26	2-WE-05
	27	2-WE-06
	28	2-WE-07
	29	2-WE-08
	30	2-WE-10
	31	2-WE-11
	32	2-WE-12
	33	2-WE-13
	34	2-WE-14
	35	2-WE-17

Tableau E.3 – Correspondances des noms des scènes enregistrées et répliquées pour l'ambiance *Rue bruyante*.

Ambiance	n°	**scène originale**
Rue bruyante	01	1-EW-01
	02	1-EW-02
	03	1-EW-06
	04	1-EW-09
	05	1-EW-12
	06	1-EW-15
	07	1-EW-18
	08	1-WE-01
	09	1-WE-02
	10	1-WE-09
	11	1-WE-10
	12	1-WE-12
	13	1-WE-18
	14	2-EW-01
	15	2-EW-02
	16	2-EW-09
	17	2-EW-10
	18	2-EW-12
	19	2-EW-14
	20	2-EW-18
	21	2-EW-01
	22	2-WE-02
	23	2-WE-09

Tableau E.4 – Correspondances des noms des scènes enregistrées et répliquées pour l'ambiance *Rue très bruyante*.

Ambiance	n°	**scène originale**
Rue très bruyante	01	1-EW-16
	02	1-WE-15
	03	1-WE-16
	04	2-EW-15
	05	2-EW-16
	06	2-WE-15
	07	2-WE-16
	08	2-WE-18

Bibliographie

[CNO, 2012] (2012). Common Noise Assessment Methods in Europe (CNOSSOS-EU) - EU Science Hub - European Commission.

[Abdallah et Plumbley, 2003] Abdallah, S. A. et Plumbley, M. D. (2003). An independent component analysis approach to automatic music transcription. *Preprints-Audio Engineering Society*.

[Ackley *et al.*, 1985] Ackley, D. H., Hinton, G. E. et Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

[Adams *et al.*, 2008] Adams, M. D., Bruce, N. S., Davies, W. J., Cain, R., Jennings, P., Carlyle, A., Cusack, P., Hume, K. et Plack, C. (2008). Soundwalking as a methodology for understanding soundscapes. *In Institute of Acoustics*, volume 30, Reading, U.K.

[Adavanne et Virtanen, 2017] Adavanne, S. et Virtanen, T. (2017). A report on sound event detection with different binaural features. Rapport technique, DCASE2017 Challenge.

[Alsina-Pagès *et al.*, 2016] Alsina-Pagès, R. M., Hernandez-Jayo, U., Alías, F. et Angulo, I. (2016). Design of a Mobile Low-Cost Sensor Network Using Urban Buses for Real-Time Ubiquitous Noise Monitoring. *Sensors*, 17(1):57.

[Aumond *et al.*, 2016] Aumond, P., Can, A., De Coensel, B., Botteldooren, D., Ribeiro, C. et Lavandier, C. (2016). Sound pleasantness evaluation of pedestrian walks in urban sound environments. *In 22nd International Congress on Acoustics (ICA 2016)*.

[Aumond *et al.*, 2017a] Aumond, P., Can, A., De Coensel, B., Botteldooren, D., Ribeiro, C. et Lavandier, C. (2017a). Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context. *Acta Acustica united with Acustica*, 103(3):430–443.

[Aumond *et al.*, 2018a] Aumond, P., Can, A., Mallet, V., De Coensel, B., Ribeiro, C., Botteldooren, D. et Lavandier, C. (2018a). Kriging-based spatial interpolation from measurements for sound level mapping in urban areas. *The Journal of the Acoustical Society of America*, 143(5):2847–2857.

[Aumond *et al.*, 2018b] Aumond, P., Jacquesson, L. et Can, A. (2018b). Probabilistic modeling framework for multisource sound mapping. *Applied Acoustics*, 139:34–43.

[Aumond *et al.*, 2017b] Aumond, P., Lavandier, C., Ribeiro, C. *et al.* (2017b). A study of the accuracy of mobile technology for measuring urban noise pollution in large scale participatory sensing campaigns. *Applied Acoustics*, 117:219–226.

[Babisch, 2008] Babisch, W. (2008). Road traffic noise and cardiovascular risk. *Noise and Health*, 10(38):27.

[Babisch *et al.*, 2005] Babisch, W., Beule, B., Schust, M. *et al.* (2005). Traffic noise and risk of myocardial infarction. *Epidemiology*, 16(1):33–40.

[Banerjee *et al.*, 2005] Banerjee, A., Merugu, S., Dhillon, I. S. et Ghosh, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749.

[Barker *et al.*, 2015] Barker, J., Marxer, R., Vincent, E. et Watanabe, S. (2015). The third 'chime'speech separation and recognition challenge : Dataset, task and baselines. *In Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 504–511. IEEE.

[Bellucci et Zambon, 2017] Bellucci, P.and Peruzzi, L. et Zambon, G. (2017). LIFE DYNAMAP project : The case study of Rome. *Applied Acoustics*, Part B(117):193–206.

[Benetos *et al.*, 2006] Benetos, E., Kotti, M. et Kotropoulos, C. (2006). Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. *In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE.

[Berry et Browne, 2005] Berry, M. et Browne, M. (2005). Email Surveillance Using Non-negative Matrix Factorization. *Computational & Mathematical Organization Theory*, 11(3): 249–264.

[Berry *et al.*, 2007] Berry, M., Browne, M., Langville, A., Pauca, V. et Plemmons, R. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173.

[Bertin, 2009] Bertin, N. (2009). *Les factorisations en matrices non-négatives : approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique*. Paris, Télécom ParisTech.

[Bertin *et al.*, 2010] Bertin, N., Badeau, R. et Vincent, E. (2010). Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):538–549.

[Bertin *et al.*, 2009] Bertin, N., Fevotte, C. et Badeau, R. (2009). A tempering approach for Itakura-Saïto non-negative matrix factorization. with application to music transcription. *In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1545–1548. IEEE.

[Bofill et Zibulevsky, 2000] Bofill, P. et Zibulevsky, M. (2000). Blind separation of more sources than mixtures using sparsity of their short-time fourier transform. *In Proc. ica*, volume 2000, pages 87–92.

[Bregman, 1990] Bregman, A. S. (1990). *Auditory Scene Analysis : The Perceptual Organization of Sound*. MIT Press.

[Bregman, 1967] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217.

[Brocolini *et al.*, 2013a] Brocolini, L., Lavandier, C., Quoy, M. et Ribeiro, C. (2013a). Measurements of acoustic environments for urban soundscapes : Choice of homogeneous periods, optimization of durations, and selection of indicators. *The Journal of the Acoustical Society of America*, 134(1):813–821.

[Brocolini *et al.*, 2013b] Brocolini, L., Lavandier, C., Quoy, M. et Ribeiro, C. (2013b). Measurements of acoustic environments for urban soundscapes : Choice of homogeneous periods, optimization of durations, and selection of indicators. *The Journal of the Acoustical Society of America*, 134(1):813–821.

[Brown, 2012] Brown, A. L. (2012). A review of progress in soundscapes and an approach to soundscape planning. *Int. J. Acoust. Vib*, 17(2):73–81.

[Brown et Cooke, 1994] Brown, G. J. et Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336.

[Brown et Wang, 2005] Brown, G. J. et Wang, D. (2005). Separation of speech by computational auditory scene analysis. *In Speech enhancement*, pages 371–402. Springer.

[Bruce *et al.*, 2009] Bruce, N., Davies, W. et Adams, M. (2009). Development of a soundscape simulator tool. *In Proceedings of the INTERNOISE Congress*, Ottawa, Canada.

[Cain *et al.*, 2013] Cain, R., Jennings, P. et Poxon, J. (2013). The development and application of the emotional dimensions of a soundscape. *Applied Acoustics*, 74(2):232–239.

[Cakir *et al.*, 2015] Cakir, E., Heittola, T., Huttunen, H. et Virtanen, T. (2015). Polyphonic sound event detection using multi label deep neural networks. *In Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–7. IEEE.

[Can *et al.*, 2014] Can, A., Dekoninck, L. et Botteldooren, D. (2014). Measurement network for urban noise assessment : Comparison of mobile measurements and spatial interpolation approaches. *Applied acoustics*, 83:32–39.

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

[Can *et al.*, 2010] Can, A., Leclercq, L., Lelong, J. et Botteldooren, D. (2010). Traffic noise spectrum analysis : Dynamic modeling vs. experimental observations. *Applied Acoustics*, 71(8):764–770.

[Cañadas-Quesada *et al.*, 2016] Cañadas-Quesada, F. J., Vera-Candeas, P., Martinez-Munoz, D., Ruiz-Reyes, N., Carabias-Orti, J. J. et Cabanas-Molero, P. (2016).

Constrained non-negative matrix factorization for score-informed piano music restoration. *Digital Signal Processing*, 50:240–257.

[Cardoso, 1998] Cardoso, J. F. (1998). Blind signal separation : statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025.

[Cazau et Nuel, 2017] Cazau, D. et Nuel, G. (2017). Understanding the Probabilistic Latent Component Analysis Framework. *ArXiv e-prints*, 1703:arXiv :1703.05208.

[Chachada et Kuo, 2014] Chachada, S. et Kuo, C.-C. J. (2014). Environmental sound recognition : A survey. *APSIPA Transactions on Signal and Information Processing*, 3.

[Chen *et al.*,] Chen, Z., Cichocki, A. et Rutkowski, T. Constrained non-Negative Matrix Factorization Method for EEG Analysis in Early Detection of Alzheimer Disease. *In 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, volume 5, pages V–V.

[Chiappetta *et al.*, 2004] Chiappetta, P., Roubaud, M.-C. et Torrésani, B. (2004). Blind source separation and the analysis of microarray data. *Journal of Computational Biology*, 11(6):1090–1109.

[Chourabi *et al.*, 2012] Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J. R., Mellouli, S., Nahon, K., Pardo, T. A. et Scholl, H. J. (2012). Understanding smart cities : An integrative framework. *In System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 2289–2297. IEEE.

[Cichocki *et al.*, 2011] Cichocki, A., Cruces, S. et Amari, S. (2011). Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization. *Entropy*, 13(1):134–170.

[Cichocki et Zdunek, 2007] Cichocki, A. et Zdunek, R. (2007). Regularized Alternating Least Squares Algorithms for Non-negative Matrix/Tensor Factorization. *In Advances in Neural Networks – ISNN 2007*, Lecture Notes in Computer Science, pages 793–802. Springer, Berlin, Heidelberg.

[Cichocki *et al.*, 2006a] Cichocki, A., Zdunek, R. et Amari, S. (2006a). Csiszár's Divergences for Non-negative Matrix Factorization : Family of New Algorithms. *In* Rosca, J., Erdogmus, D., Príncipe, J. C. et Haykin, S., éditeurs : *Independent Component Analysis and Blind Signal Separation*, numéro 3889 de Lecture Notes in Computer Science, pages 32–39. Springer Berlin Heidelberg. DOI : 10.1007/11679363_5.

[Cichocki *et al.*, 2006b] Cichocki, A., Zdunek, R. et Amari, S.-I. (2006b). Csiszar's divergences for non-negative matrix factorization : Family of new algorithms. *In International Conference on Independent Component Analysis and Signal Separation*, pages 32–39. Springer.

[Cichocki *et al.*, 2006c] Cichocki, A., Zdunek, R. et Amari, S.-I. (2006c). New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation. *In 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, volume 5, pages V–V.

[Comon, 1994] Comon, P. (1994). Independent component analysis, A new concept ? *Signal Processing*, 36(3):287–314.

[Csiszár *et al.*, 2004] Csiszár, I., Shields, P. *et al.* (2004). Information theory and statistics : A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528.

[Davies *et al.*, 2014] Davies, W. J., Bruce, N. S. et Murphy, J. E. (2014). Soundscape reproduction and synthesis. *Acta Acustica United with Acustica*, 100(2):285–292.

[Defréville *et al.*, 2006] Defréville, B., Pachet, F., Rosin, C. et Roy, P. (2006). Automatic Recognition of Urban Sound Sources. Audio Engineering Society.

[Delaitre *et al.*, 2014] Delaitre, P., Lavandier, C., Ribeiro, C., Quoy, M., D'Hondt, E., Gonzalez Boix, E. et Kambona, K. (2014). Influence of loudness of noise events on perceived sound quality in urban context. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 249, pages 1505–1514. Institute of Noise Control Engineering.

[Delorme *et al.*, 2007] Delorme, A., Sejnowski, T. et Makeig, S. (2007). Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. *Neuroimage*, 34(4):1443–1449.

[Dessein *et al.*, 2013] Dessein, A., Cont, A. et Lemaitre, G. (2013). Real-time detection of overlapping sound events with non-negative matrix factorization. *In Matrix Information Geometry*, pages 341–371. Springer.

[Dhillon et Sra, 2005] Dhillon, I. S. et Sra, S. (2005). Generalized nonnegative matrix approximations with Bregman divergences. *In In : Neural Information Proc. Systems*, pages 283–290.

[Donoho et Johnstone, 1994] Donoho, D. L. et Johnstone, I. M. (1994). Threshold selection for wavelet shrinkage of noisy data. *In Engineering in Medicine and Biology Society, 1994. Engineering Advances : New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, volume 1, pages A24–A25. IEEE.

[Duan *et al.*, 2012] Duan, Z., Mysore, G. J. et Smaragdis, P. (2012). Online plca for real-time semi-supervised source separation. *In International Conference on Latent Variable Analysis and Signal Separation*, pages 34–41. Springer.

[Dufaux *et al.*, 2000] Dufaux, A., Besacier, L., Ansorge, M. et Pellandini, F. (2000). Automatic sound detection and recognition for noisy environment. *In Signal Processing Conference, 2000 10th European*, pages 1–4. IEEE.

[Dutilleux, 2012] Dutilleux, G. (2012). Anthropogenic outdoor sound and wildlife : it's not just bioacoustics ! *In* d'Acoustique, S. F., éditeur : *Acoustics 2012*, Nantes, France.

[Eggert et Körner, 2004] Eggert, J. et Körner, E. (2004). Sparse coding and nmf. *In Conference : Neural Networks, 2004. Proceedings. 2004 IEEE International*, volume 4, pages 2529–2533.

[El Ayadi *et al.*, 2011] El Ayadi, M., Kamel, M. S. et Karray, F. (2011). Survey on speech emotion recognition : Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

[Ellis, 1999] Ellis, D. P. (1999). Using knowledge to organize sound : The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures1. *Speech Communication*, 27(3-4):281–298.

[Eronen *et al.*, 2006] Eronen, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., Lorho, G. et Huopaniemi, J. (2006). Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):321–329.

[Essid et Févotte, 2013] Essid, S. et Févotte, C. (2013). Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Transactions on Multimedia*, 15(2):415–425.

[European Environment Agency, 2014] European Environment Agency (2014). Noise in Europe.

[EUROPEENS, 2016] EUROPEENS, T. F. e. E. (2016). Analyse bibliographique des travaux français et europeens : Le coût social des pollutions sonores. (59 p.).

[Févotte, 2011] Févotte, C. (2011). Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorization. *In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 1980–1983. IEEE.

[Févotte *et al.*, 2009] Févotte, C., Bertin, N. et Durrieu, J. (2009). Nonnegative matrix factorization with the Itakura-Saïto divergence : with application to music analysis. *Neural Computation*, 21(3):793–830.

[Févotte et Idier, 2011] Févotte, C. et Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456.

[Févotte *et al.*, 2018] Févotte, C., Vincent, E. et Ozerov, A. (2018). Single-channel audio source separation with nmf : divergences, constraints and algorithms. *In Audio Source Separation*, pages 1–24. Springer.

[Finney et Janer, 2010] Finney, N. et Janer, J. (2010). Soundscape generation for virtual environments using community-provided audio databases. *In W3C Workshop : Augmented Reality on the Web*. Barcelona.

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

[Fornasier et Rauhut, 2008] Fornasier, M. et Rauhut, H. (2008). Iterative thresholding algorithms. *Applied and Computational Harmonic Analysis*, 25(2):187–208.

[Forssén *et al.*, 2009] Forssén, J., Kaczmarek, T., Alvarsson, J., Lundén, P. et Nilsson, M. E. (2009). Auralization of traffic noise within the listen project—preliminary results for passenger car pass-by. *Euronoise 2009*.

[Fortin *et al.*, 2012] Fortin, N., Bocher, E., Picaut, J., Petit, G. et Dutilleux, G. (2012). An opensource tool to build urban noise maps in a GIS. *In Open Source Geospatial Research and Education Symposium (OGRS)*, pages 9p, cartes, Yverdon-Les-Bains, Switzerland.

[Francis *et al.*, 2009] Francis, C. D., Ortega, C. P. et Cruz, A. (2009). Noise pollution changes avian communities and species interactions. *Current biology*, 19(16):1415–1419.

[Févotte, 2011] Févotte, C. (2011). Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. *In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 1980–1983. IEEE.

[Gaja *et al.*, 2003] Gaja, E., Gimenez, A., Sancho, S. et Reig, A. (2003). Sampling techniques for the estimation of the annual equivalent noise level under urban traffic conditions. *Applied Acoustics*, 64(1):43–53.

[Gannot *et al.*, 2017] Gannot, S., Vincent, E., Markovich-Golan, S., Ozerov, A., Gannot, S., Vincent, E., Markovich-Golan, S. et Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(4):692–730.

[Gao et Church, 2005] Gao, Y. et Church, G. (2005). Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21(21):3970–3975.

[Garg et Maji, 2014] Garg, N. et Maji, S. (2014). A Critical Review of Principal Traffic Noise Models : Strategies and Implications. *Environmental Impact Assessment Review*, 46.

[Gaussier et Goutte, 2005] Gaussier, E. et Goutte, C. (2005). Relation Between PLSA and NMF and Implications. SIGIR '05, pages 601–602, New York, NY, USA. ACM.

[Geiger et Helwani, 2015] Geiger, J. T. et Helwani, K. (2015). Improving event detection for audio surveillance using gabor filterbank features. *In Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 714–718. IEEE.

[Gemmeke *et al.*, 2013] Gemmeke, J. F., Vuegen, L., Karsmakers, P., Vanrumste, B. *et al.* (2013). An exemplar-based nmf approach to audio event detection. *In Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pages 1–4. IEEE.

[Guastavino *et al.*, 2005] Guastavino, C., Katz, B. F. G., Polack, J., Levitin, D. J. et Dubois, D. (2005). Ecological validity of soundscape reproduction. *Acta Acustica united with Acustica*, 91(2):333–341.

[Guedes *et al.*, 2011] Guedes, I. C. M., Bertoli, S. R. et Zannin, P. H. (2011). Influence of urban shapes on environmental noise : a case study in aracaju—brazil. *Science of the Total Environment*, 412:66–76.

[Guillamet *et al.*, 2003] Guillamet, D., Vitrià, J. et Schiele, B. (2003). Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recognition Letters*, 24(14):2447–2454.

[Guillaume *et al.*, 2016] Guillaume, G., Can, A., Petit, G., Fortin, N., Palominos, S., Gauvreau, B., Bocher, E. et Picaut, J. (2016). Noise mapping based on participative measurements. *Noise Mapp*, 3:140–156.

[Harlow et Wang, 2001] Harlow, C. et Wang, Y. (2001). Automated accident detection system. *Transportation Research Record : Journal of the Transportation Research Board*, (1746):90–93.

[Hawkins, 2006] Hawkins, D. (2006). Clustering scotch whiskies using non-negative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association*, (14):11–13.

[Hayne *et al.*, 2011] Hayne, M., Taylor, J., Rumble, R. et Mee, D. (2011). Prediction of noise from small to medium sized crowds. *Proceedings of Acoustics 2011*.

[Heittola *et al.*, 2011] Heittola, T., Mesaros, A., Virtanen, T. et Eronen, A. (2011). Sound event detection in multisource environments using source separation. *In in Workshop on Machine Listening in Multisource Environments, CHiME2011*.

[Helen et Virtanen, 2005] Helen, M. et Virtanen, T. (2005). Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. *In Signal Processing Conference, 2005 13th European*, pages 1–4. IEEE.

[Hennequin *et al.*, 2011a] Hennequin, R., Badeau, R. et David, B. (2011a). Scale-invariant probabilistic latent component analysis. *In 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 129–132.

[Hennequin *et al.*, 2011b] Hennequin, R., David, B. et Badeau, R. (2011b). Beta-Divergence as a Subclass of Bregman Divergence. *IEEE Signal Processing Letters*, 18(2):83–86.

[Hofmann, 2001] Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196.

[Hong et Jeon, 2013] Hong, J. Y. et Jeon, J. Y. (2013). Designing sound and visual components for enhancement of urban soundscapes. *The Journal of the Acoustical Society of America*, 134(3):2026–2036.

[Hoyer, 2004] Hoyer, P. (2004). Non-negative matrix factorization with sparseness constraints. *Jour. of*, pages 1457–1469.

[Hsieh *et al.*, 2009] Hsieh, H.-L., Chien, J.-T., Shinoda, K. et Furui, S. (2009). Independent component analysis for noisy speech recognition. *In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4369–4372. IEEE.

[Hu et Wang, 2006] Hu, G. et Wang, D. (2006). An auditory scene analysis approach to monaural speech segregation. *Topics in Acoustic Echo and Noise Control : Selected Methods for the Cancellation of Acoustical Echoes, the Reduction of Background Noise, and Speech Processing*, page 485.

[Huang *et al.*, 2014] Huang, P.-S., Kim, M., Hasegawa-Johnson, M. et Smaragdis, P. (2014). Singing-voice separation from monaural recordings using deep recurrent neural networks. *In ISMIR*, pages 477–482.

[Hurmalainen *et al.*, 2012] Hurmalainen, A., Gemmeke, J. F. et Virtanen, T. (2012). Detection, separation and recognition of speech from continuous signals using spectral factorisation.

In Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, pages 2649–2653. IEEE.

[Hyvärinen, 1999] Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634.

[Hyvärinen *et al.*, 2004] Hyvärinen, A., Karhunen, J. et Oja, E. (2004). *Independent component analysis*, volume 46. John Wiley & Sons.

[Hyvärinen, 1997] Hyvärinen, A. (1997). Independent component analysis by minimization of mutual information.

[Innami et Kasai, 2012] Innami, S. et Kasai, H. (2012). Nmf-based environmental sound source separation using time-variant gain features. *Computers & Mathematics with Applications*, 64(5):1333–1342.

[Ising *et al.*, 1980] Ising, H., Dienel, D., Günther, T. et Markert, B. (1980). Health effects of traffic noise. *International Archives of Occupational and Environmental Health*, 47(2):179–190.

[Itakura, 1968] Itakura, F. (1968). Analysis synthesis telephony based on the maximum likelihood method. *In The 6th international congress on acoustics, 1968*, pages 280–292.

[Jagniatinskis et Fiks, 2014] Jagniatinskis, A. et Fiks, B. (2014). Assessment of environmental noise from long-term window microphone measurements. *Applied Acoustics*, 76:377–385.

[Jonasson *et al.*, 2004] Jonasson, H., Sandberg, U., Blokland, G. v., Ejsmont, J., Watts, G. et Luminari, M. (2004). Source modelling of road vehicles.

[Jorgensen, 1987] Jorgensen, B. (1987). Exponential Dispersion Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2):127–162.

[Jourjine *et al.*, 2000] Jourjine, A., Rickard, S. et Yilmaz, O. (2000). Blind separation of disjoint orthogonal signals : Demixing n sources from 2 mixtures. *In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 5, pages 2985–2988. IEEE.

[Jung *et al.*, 2000] Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V. et Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178.

[Jutten et Herault, 1991] Jutten, C. et Herault, J. (1991). Blind separation of sources, part i : An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10.

[Kanjo, 2010] Kanjo, E. (2010). NoiseSPY : A Real-Time Mobile Phone Platform for Urban Noise Monitoring and Mapping. *Mobile Networks and Applications*, 15(4):562–574.

[Kephalopoulos *et al.*, 2012] Kephalopoulos, ., PAVIOTTI, M. et Anfosso-Lédée, F. (2012). *Common noise assessment methods in Europe (CNOSSOS-EU)*. PUBLICATIONS OFFICE OF THE EUROPEAN UNION.

[King *et al.*, 2011] King, E. A., Murphy, E. et Rice, H. J. (2011). Implementation of the EU environmental noise directive : lessons from the first phase of strategic noise mapping and action planning in Ireland. *Journal of Environmental Management*, 92(3):756–764.

[Kitamura *et al.*, 2014] Kitamura, D., Saruwatari, H., Yagi, K., Shikano, K., Takahashi, Y. et Kondo, K. (2014). Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 97(5):1113–1118.

[Kivinen et Warmuth, 1994] Kivinen, J. et Warmuth, M. (1994). Exponentiated gradient versus gradient descent for linear predictors. Rapport technique, University of California at Santa Cruz, Santa Cruz, CA, USA.

[Kliučininkas et Šaliūnas, 2006] Kliučininkas, L. et Šaliūnas, D. (2006). Noise mapping for the management of urban traffic flows. *Mechanics*, 59(3):61–66.

[Komatsu *et al.*, 2016] Komatsu, T., Toizumi, T., Kondo, R. et Senda, Y. (2016). Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries. Rapport technique, DCASE2016 Challenge.

[Kominek et Black, 2004] Kominek, J. et Black, A. W. (2004). The cmu arctic speech databases. *In Fifth ISCA Workshop on Speech Synthesis*.

[Kompass, 2007] Kompass, R. (2007). A Generalized Divergence Measure for Nonnegative Matrix Factorization. *Neural Computation*, 19(3):780–791.

[Kuhn, 1976] Kuhn, H. W. (1976). Nonlinear programming : a historical view. *SIAM-AMS proceedings*, (9):1–22.

[Kumar *et al.*, 2016] Kumar, A., Elizalde, B. et Raj, B. (2016). Audio content based geotagging in multimedia. *arXiv preprint arXiv :1606.02816*.

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

[Lafay *et al.*, 2015] Lafay, G., Lagrange, M., Petiot, J.-F., Rossignol, M. et Misdariis, N. (2015). Approaching mental representations of urban soundscape using auditory scenes simulation. working paper or preprint.

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

[Lagrange *et al.*, 2015] Lagrange, M., Lafay, G., Rossignol, M., Benetos, E. et Roebel, A. (2015). An evaluation framework for event detection using a morphological model of acoustic scenes. *arXiv preprint arXiv :1502.00141*.

[Lavandier et Defréville, 2006] Lavandier, C. et Defréville, B. (2006). The contribution of sound source characteristics in the assessment of urban soundscapes. *Acta Acustica united with Acustica*, 92(6):912–921.

[Le Roux *et al.*, 2015] Le Roux, J., Weninger, F. J. et Hershey, J. R. (2015). Sparse nmf–half-baked or well done ? *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*.

[LeCun *et al.*, 2015] LeCun, Y., Bengio, Y. et Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

[Lee et Seung, 2000] Lee, D. et Seung, H. (2000). Algorithms for Non-negative Matrix Factorization. *In In NIPS*, pages 556–562. MIT Press.

[Lee et Seung, 1999] Lee, D. D. et Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

[Lee *et al.*, 2010] Lee, H., Yoo, J. et Choi, S. (2010). Semi-Supervised Nonnegative Matrix Factorization. *IEEE Signal Processing Letters*, 17(1):4–7.

[Lee *et al.*, 2000] Lee, T.-W., Girolami, M., Bell, A. J. et Sejnowski, T. J. (2000). A unifying information-theoretic framework for independent component analysis. *Computers & Mathematics with Applications*, 39(11):1–21.

[Lee *et al.*, 1999] Lee, T.-W., Girolami, M. et Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation*, 11(2):417–441.

[Lefevre *et al.*, 2012] Lefevre, A., Bach, F. et Févotte, C. (2012). Semi-supervised {NMF} with time-frequency annotations for single-channel source separation. *In ISMIR 2012 : 13th International Society for Music Information Retrieval Conference*.

[Leiba *et al.*, 2017] Leiba, R., Ollivier, F., Marchal, J., Misdariis, N., Marchiano, R. *et al.* (2017). Large array of microphones for the automatic recognition of acoustic sources in urban environment. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 255, pages 2662–2670. Institute of Noise Control Engineering.

[Lettvin *et al.*, 1959] Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. et Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11):1940–1951.

[Li *et al.*, 2001] Li, S. Z., Hou, X. W., Zhang, H. J. et Cheng, Q. S. (2001). Learning spatially localized, parts-based representation. *In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

[Li et Ding, 2006] Li, T. et Ding, C. (2006). The relationships among various nonnegative matrix factorization methods for clustering. *In Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 362–371. IEEE.

[Lin, 2007] Lin, C. J. (2007). Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*, 19(10):2756–2779.

[Liu *et al.*, 2014] Liu, J., Kang, J., Behm, H. et Luo, T. (2014). Effects of landscape on soundscape perception : Soundwalks in city parks. *Landscape and Urban Planning*, 123:30–40.

[Lombard *et al.*, 2011] Lombard, A., Zheng, Y., Buchner, H. et Kellermann, W. (2011). Tdoa estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1490–1503.

[Luo *et al.*, 2014] Luo, X., Zhou, M., Xia, Y. et Zhu, Q. (2014). An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284.

[Lv *et al.*, 2015] Lv, Y., Duan, Y., Kang, W. *et al.* (2015). Traffic flow prediction with big data : a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873.

[Makarewicz et Galuszka, 2011] Makarewicz, R. et Galuszka, M. (2011). Empirical revision of noise mapping. *Applied Acoustics*, 72(8):578–581.

[Makeig *et al.*, 1996] Makeig, S., Bell, A. J., Jung, T.-P. et Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. *In Advances in neural information processing systems*, pages 145–151.

[Manvell *et al.*, 2004] Manvell, D., Ballarin Marcos, L., Stapelfeldt, H. et Sanz, R. (2004). Sadmam-combining measurements and calculations to map noise in madrid. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2004, pages 1998–2005. Institute of Noise Control Engineering.

[Melia et Rickard, 2007] Melia, T. et Rickard, S. (2007). Underdetermined blind source separation in echoic environments using desprit. *EURASIP Journal on Applied Signal Processing*, 2007(1):90–90.

[Mesaros *et al.*, 2015] Mesaros, A., Heittola, T., Dikmen, O. et Virtanen, T. (2015). Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations. *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155.

[Mesaros *et al.*, 2017] Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B. et Virtanen, T. (2017). Dcase 2017 challenge setup : Tasks, datasets and baseline system. *In DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*.

[Mesaros *et al.*, 2010] Mesaros, A., Heittola, T., Eronen, A. et Virtanen, T. (2010). Acoustic event detection in real life recordings. *In Signal Processing Conference, 2010 18th European*, pages 1267–1271. IEEE.

[Mietlicki *et al.*, 2012] Mietlicki, C., Mietlicki, F. et Sineau, M. (2012). An innovative approach for long-term environmental noise measurement : Rumeur network. *In InterNoise and NoiseCon Congress and Conference Proceedings*, pages 7119–7130. Institute of Noise Control Engineering.

[Mioduszewski *et al.*, 2011] Mioduszewski, P., Ejsmont, J. A., Grabowski, J. et Karpinski, D. (2011). Noise map validation by continuous noise monitoring. *Applied Acoustics*, 72(8):582–589.

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

[Monga et Mhcak, 2007] Monga, V. et Mhcak, M. K. (2007). Robust and Secure Image Hashing via Non-Negative Matrix Factorizations. *IEEE Transactions on Information Forensics and Security*, 3-1(2):376–390.

[Morillas et Gajardo, 2014] Morillas, J. B. et Gajardo, C. P. (2014). Uncertainty evaluation of continuous noise sampling. *Applied Acoustics*, 75:27–36.

[Murphy et King, 2011] Murphy, E. et King, E. A. (2011). Scenario analysis and noise action planning : Modelling the impact of mitigation measures on population exposure. *Applied Acoustics*, 72(8):487–494.

[Murphy *et al.*, 2006] Murphy, E., Rice, H. J. et Meskell, C. (2006). Environmental noise prediction, noise mapping and GIS integration : the case of inner Dublin, Ireland. East-European Acoustical Association.

[Mydlarz *et al.*, 2015] Mydlarz, C., Shamoon, C., Baglione, M. et Pimpinella, M. (2015). The design and calibration of low cost urban acoustic sensing devices. *In Proceedings of the EuroNoise*.

[Mydlarz *et al.*, 2017] Mydlarz, C., Shamoon, C., Bello, J. P. *et al.* (2017). Noise monitoring and enforcement in new york city using a remote acoustic sensor network. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 255, pages 5509–5520. Institute of Noise Control Engineering.

[Mysore et Smaragdis, 2011] Mysore, G. J. et Smaragdis, P. (2011). A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. *In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 17–20. IEEE.

[Naik et Kumar, 2011] Naik, G. R. et Kumar, D. K. (2011). An overview of independent component analysis and its applications. *Informatica*, 35(1).

[Nantes Métropole, 2017] Nantes Métropole (2017). Les cartes des bruits. http ://www.nantesmetropole.fr/la-communaute-urbaine/competences/les-cartes-des-bruits-27856.kjsp. visité le 24/08/2017.

[Nelken et De Cheveigné, 2013] Nelken, I. et De Cheveigné, A. (2013). An ear for statistics. *Nature neuroscience*, 16(4):381.

[Nemeth *et al.*, 2013] Nemeth, E., Pieretti, N., Zollinger, S. A., Geberzahn, N., Partecke, J., Miranda, A. C. et Brumm, H. (2013). Bird song and anthropogenic noise : vocal constraints may explain why birds sing higher-frequency songs in cities. *Proc. R. Soc. B*, 280(1754):20122798.

[Nickalls, 1993] Nickalls, R. W. (1993). A new approach to solving the cubic : Cardan's solution revealed. *The Mathematical Gazette*, 77(480):354–359.

[Ntalampiras, 2014] Ntalampiras, S. (2014). Universal background modeling for acoustic surveillance of urban traffic. *Digital Signal Processing*, 31:69–78.

[Ntalampiras *et al.*, 2011] Ntalampiras, S., Potamitis, I. et Fakotakis, N. (2011). Probabilistic novelty detection for acoustic surveillance under real-world conditions. *IEEE Transactions on Multimedia*, 13(4):713–719.

[Nugraha *et al.*, 2016] Nugraha, A. A., Liutkus, A. et Vincent, E. (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(9):1652–1664.

[Nuzillard et Bijaoui, 2000] Nuzillard, D. et Bijaoui, A. (2000). Blind source separation and analysis of multispectral astronomical images. *Astronomy and Astrophysics Supplement Series*, 147(1):129–138.

[Paatero et Tapper, 1994] Paatero, P. et Tapper, U. (1994). Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

[Pagès et Périnel, 2007] Pagès, J. et Périnel, E. (2007). Blocs incomplets équilibrés versus plans optimaux. *Journal de la Société Française de Statistique*, 148(2):99–112.

[Panagakis *et al.*, 2008] Panagakis, I., Benetos, E. et Kotropoulos, C. (2008). Music genre classification : A multilinear approach. *In ISMIR*, pages 583–588.

[Parlement Européen, 2002] Parlement Européen (2002). Directive 2002/49/ce du parlement européen et du conseil du 25 juin 2002 relative à l'évaluation et à la gestion du bruit dans l'environnement - déclaration de la commission au sein du comité de conciliation concernant la directive relative à l'évaluation et à la gestion du bruit ambiant. *Journal officiel n° L 189 du 18/07/2002*, pages 12 – 26.

[Pascual-Montano *et al.*, 2006] Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D. et Pascual-Marqui, R. D. (2006). Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE transactions on pattern analysis and machine intelligence*, 28(3):403–415.

[Peltonen *et al.*, 2002] Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J. et Sorsa, T. (2002). Computational auditory scene recognition. *In Acoustics, speech, and signal processing (icassp), 2002 IEEE international conference on*, volume 2, pages II–1941. IEEE.

[Picaut *et al.*, 2017] Picaut, J., Can, A., Ardouin, J., Crépeaux, P., Dhorne, T., Écotière, D., Lagrange, M., Lavandier, C., Mallet, V., Mietlicki, C. *et al.* (2017). Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling. *The Journal of the Acoustical Society of America*, 141(5):3808–3808.

[Picaut *et al.*, 2005] Picaut, J., Le Pollès, T., L'Hermite, P. et Gary, V. (2005). Experimental study of sound propagation in a street. *Applied Acoustics*, 66(2):149–173.

[Piczak, 2015] Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. *In Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pages 1–6. IEEE.

[Pirrera *et al.*, 2010] Pirrera, S., De Valck, E. et Cluydts, R. (2010). Nocturnal road traffic noise : A review on its assessment and consequences on sleep and health. *Environment international*, 36(5):492–498.

[Probst *et al.*, 2011] Probst, F., Probst, W. et Huber, B. (2011). Comparison of noise calculation methods. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2011, pages 962–967. Institute of Noise Control Engineering.

[Quartieri *et al.*, 2009] Quartieri, J., Mastorakis, N. E., Iannone, G. *et al.* (2009). A review of traffic noise predictive models. *In Recent Advances in Applied and Theoretical Mechanics, 5th WSEAS International Conference on Applied and Theoretical Mechanics (MECHANICS'09) Puerto De La Cruz, Tenerife, Canary Islands, Spain December*, pages 14–16.

[Rickard, 2007] Rickard, S. (2007). The duet blind source separation algorithm. *In Blind Speech Separation*, pages 217–241. Springer.

[Rickard *et al.*, 2001] Rickard, S., Balan, R. et Rosca, J. (2001). Real-time time-frequency based blind source separation. *aje*, 2:1.

[Rigaud *et al.*, 2012] Rigaud, F., David, B. et Daudet, L. (2012). Piano sound analysis using non-negative matrix factorization with inharmonicity constraint. *In Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2462–2466. IEEE.

[Romeu *et al.*, 2011] Romeu, J., Genescà, M., Pàmies, T. et Jiménez, S. (2011). Street categorization for the estimation of day levels using short-term measurements. *Applied acoustics*, 72(8):569–577.

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

[Ruxton, 2006] Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690.

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

[Salamon et Bello, 2015] Salamon, J. et Bello, J. P. (2015). Unsupervised feature learning for urban sound classification. *In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 171–175. IEEE.

[Salamon et Bello, 2017] Salamon, J. et Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283.

[Salamon *et al.*, 2014] Salamon, J., Jacoby, C. et Bello, J. P. (2014). A dataset and taxonomy for urban sound research. *In 22st ACM International Conference on Multimedia (ACM-MM 14)*, Orlando, FL, USA.

[Särelä et Valpola, 2005] Särelä, J. et Valpola, H. (2005). Denoising source separation. *Journal of machine learning research*, 6(Mar):233–272.

[Saruwatari *et al.*, 2003] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T. et Shikano, K. (2003). Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Advances in Signal Processing*, 2003(11):569270.

[Schissler *et al.*, 2014] Schissler, C., Mehra, R. et Manocha, D. (2014). High-order diffraction and diffuse reflections for interactive sound propagation in large environments. *ACM Transactions on Graphics (TOG)*, 33(4):39.

[Schmidhuber, 2015] Schmidhuber, J. (2015). Deep learning in neural networks : An overview. *Neural networks*, 61:85–117.

[Schmidt et Olsson, 2006] Schmidt, M. N. et Olsson, R. K. (2006). Single-channel speech separation using sparse non-negative matrix factorization. *In Ninth International Conference on Spoken Language Processing*.

[SETRA, 2009a] SETRA (2009a). *Prévision du bruit routier - 1 - Calcul des émissions sonores dues au trafic routier*, volume 1. Setra édition.

[SETRA, 2009b] SETRA (2009b). *Prévision du bruit routier - 2 - Méthode de calcul de propagation du bruit incluant les effets météorologiques (NMPB 2008)*, volume 2. Setra édition.

[Shao *et al.*, 2010] Shao, Y., Srinivasan, S., Jin, Z. et Wang, D. (2010). A computational auditory scene analysis system for speech segregation and robust speech recognition. *Computer Speech & Language*, 24(1):77–93.

[Shashanka *et al.*, 2008] Shashanka, M., Raj, B. et Smaragdis, P. (2008). Probabilistic Latent Variable Models as Nonnegative Factorizations. DOI : 10.1155/2008/947438.

[Smaragdis, 2007] Smaragdis, P. (2007). Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1–12.

[Smaragdis et Brown, 2003] Smaragdis, P. et Brown, J. (2003). Non-negative matrix factorization for polyphonic music transcription. *In Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180.

[Smaragdis et Raj, 2007] Smaragdis, P. et Raj, B. (2007). Shift-invariant probabilistic latent component analysis. Rapport technique, Mitsubishi Electric Research Laboratories.

[Smaragdis *et al.*, 2007] Smaragdis, P., Raj, B. et Shashanka, M. (2007). Supervised and semi-supervised separation of sounds from single-channel mixtures. *Independent Component Analysis and Signal Separation*, pages 414–421.

[Sobieraj *et al.*, 2017] Sobieraj, I., Kong, Q. et Plumbley, M. D. (2017). Masked non-negative matrix factorization for eire detection using weakly labeled data. *In Signal Processing Conference (EUSIPCO), 2017 25th European*, pages 1769–1773. IEEE.

[Socoró *et al.*, 2017] Socoró, J. C., Alías, F. et Alsina-Pagès, R. M. (2017). An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments. *Sensors*, 17(10):2323.

[Souviraà-Labastie *et al.*, 2015] Souviraà-Labastie, N., Vincent, E. et Bimbot, F. (2015). Music separation guided by cover tracks : Designing the joint nmf model. *In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 484–488. IEEE.

[Sprechmann *et al.*, 2014] Sprechmann, P., Bronstein, A. M. et Sapiro, G. (2014). Supervised non-euclidean sparse nmf via bilevel optimization with applications to speech enhancement. *In Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, pages 11–15. IEEE.

[Steele, 2001] Steele, C. (2001). A critical review of some traffc noise prediction models. *Applied Acoustics*, 62(3):271–287.

[Stienen et Vorländer, 2015] Stienen, J. et Vorländer, M. (2015). Auralization of urban environments—concepts towards new applications. *In Proc. EuroNoise*, Maastricht, Pays-Bas.

[Stowell *et al.*, 2015] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M. et Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746.

[Torija et Ruiz, 2012] Torija, A. J. et Ruiz, D. P. (2012). Using recorded sound spectra profile as input data for real-time short-term urban road-traffic-flow estimation. *Science of the Total Environment*, 435:270–279.

[Uhle *et al.*, 2003] Uhle, C., Dittmar, C. et Sporer, T. (2003). Extraction of drum tracks from polyphonic music using independent subspace analysis. *In Proc. ICA*, pages 843–847.

[Valle *et al.*, 2009] Valle, A., Schirosa, M. et Lombardo, V. (2009). A framework for soundscape analysis and re-synthesis. *Proceedings of the SMC*, pages 13–18.

[Van Leeuwen et Van Banda, 2015] Van Leeuwen, H. et Van Banda, S. (2015). Noise mapping - State of the art - Is it just as simple as it looks ? *EuroNoise*.

[van Leeuwen, 2000] van Leeuwen, H. J. A. (2000). Railway noise prediction models : A comparison. *Journal of Sound and Vibration*, 231(3):975–987.

[Van Renterghem *et al.*, 2010] Van Renterghem, T., Thomas, P., Botteldooren, D. *et al.* (2010). The use of cheap microphones in extensive outdoor noise monitoring networks. *In Noise in the Built Environment : a joint conference organised by the institute of acoustics & the Belgium acoustical association*, volume 32, pages 374–377. Institute of Acoustics.

[Ventura *et al.*, 2017] Ventura, R., Mallet, V., Issarny, V. *et al.* (2017). Estimation of urban noise with the assimilation of observations crowdsensed by the mobile application ambiciti. *In INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 255, pages 5444–5451. Institute of Noise Control Engineering.

[Vincent, 2006] Vincent, E. (2006). Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):91–98.

[Vincent *et al.*, 2008] Vincent, E., Bertin, N. et Badeau, R. (2008). Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. *In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 109–112. IEEE.

[Vincent *et al.*, 2010] Vincent, E., Bertin, N. et Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537.

[Vincent *et al.*, 2006] Vincent, E., Gribonval, R. et Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469.

[Virtanen, 2003] Virtanen, T. (2003). Sound Source Separation Using Sparse Coding with Temporal Continuity Objective. *International Computer Music Conference Proceedings*, 2003.

[Virtanen, 2004] Virtanen, T. (2004). Separation of sound sources by convolutive sparse coding. *In ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*.

[Virtanen, 2007] Virtanen, T. (2007). Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074.

[Vorländer, 2007] Vorländer, M. (2007). *Auralization : fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media.

[Wang et Brown, 2006] Wang, D. et Brown, G. J. (2006). *Computational auditory scene analysis : Principles, algorithms, and applications*. Wiley-IEEE press.

[Wang *et al.*, 2016] Wang, D., Gao, X. et Wang, X. (2016). Semi-supervised nonnegative matrix factorization via constraint propagation. *IEEE transactions on cybernetics*, 46(1):233–244.

[Watts *et al.*, 2009] Watts, G. R., Pheasant, R. J., Horoshenkov, K. V. et Ragonesi, L. (2009). Measurement and subjective assessment of water generated sounds. *Acta Acustica united with Acustica*, 95(6):1032–1039.

[Wei *et al.*, 2016] Wei, W., Renterghem, T. V., Coensel, B. D. et Botteldooren, D. (2016). Dynamic noise mapping : A map-based interpolation between noise measurements with high temporal resolution. *Applied Acoustics*, Complete(101):127–140.

[Weninger *et al.*, 2012] Weninger, F., Feliu, J. et Schuller, B. (2012). Supervised and semi-supervised suppression of background music in monaural speech recordings. *In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 61–64. IEEE.

[WHO, 2017] WHO, W. (2017). Noise - data and statistics. http ://www.euro.who.int/en/health-topics/environment-and-health/noise/data-and-statistics. visité le 24/08/2017.

[Wilson *et al.*, 2008a] Wilson, K. W., Raj, B., Smaragdis, P. et Divakaran, A. (2008a). Speech denoising using nonnegative matrix factorization with priors. *In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4029–4032. IEEE.

[Wilson *et al.*, 2008b] Wilson, K. W., Raj, B., Smaragdis, P. et Divakaran, A. (2008b). Speech denoising using nonnegative matrix factorization with priors. *In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4029–4032.

[Wu *et al.*, 2008] Wu, L., Mallet, V., Bocquet, M. et Sportisse, B. (2008). A comparison study of data assimilation algorithms for ozone forecasts. *Journal of Geophysical Research : Atmospheres*, 113(D20).

[Xavier *et al.*, 2016] Xavier, S., Claudi, S. J., Francesc, A., Patrizia, B. et al. (2016). DYNAMAP – Development of low cost sensors networks for real time noise mapping. *Noise Mapping*, 3(1).

[Xu *et al.*, 2003] Xu, W., Liu, X. et Gong, Y. (2003). Document Clustering Based on Non-negative Matrix Factorization. *In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 267–273, New York, NY, USA. ACM.

[Yagi *et al.*, 2012] Yagi, K., Takahashi, Y., Saruwatari, H., Shikano, K. et Kondo, K. (2012). Music signal separation by orthogonality and maximum-distance constrained non-negative matrix factorization with target signal information. *In Audio Engineering Society Conference : 45th International Conference : Applications of Time-Frequency Processing in Audio*. Audio Engineering Society.

[Yilmaz et Rickard, 2004] Yilmaz, O. et Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 52(7):1830–1847.

[Yilmaz et Cemgil, 2012] Yilmaz, Y. K. et Cemgil, A. T. (2012). Alpha/beta divergences and Tweedie models. *arXiv preprint arXiv :1209.4280*.

[Zambon *et al.*, 2017] Zambon, G., Benocci, R., Bisceglie, A. *et al.* (2017). The life dynamap project : Towards a procedure for dynamic noise mapping in urban areas. *Applied Acoustics*, 124:52–60.

[Zanella *et al.*, 2014] Zanella, A., Bui, N., Castellani, A. *et al.* (2014). Internet of things for smart cities. *IEEE Internet of Things journal*, 1(1):22–32.

[Zannin *et al.*, 2013] Zannin, P. H., Engel, M. S., Fiedler, P. E. et Bunn, F. (2013). Characterization of environmental noise based on noise measurements, noise mapping and interviews : A case study at a university campus in Brazil. *Cities*, 31:317–327.

[Zannin *et al.*, 2002] Zannin, P. H. T., Diniz, F. B. et Barbosa, W. A. (2002). Environmental noise pollution in the city of curitiba, brazil. *Applied Acoustics*, 63(4):351–358.

[Zaporozhets et Tokarev, 1998] Zaporozhets, O. I. et Tokarev, V. I. (1998). Aircraft noise modelling for environmental assessment around airports. *Applied Acoustics*, 55(2):99–127.

[Zdunek et Cichocki, 2006] Zdunek, R. et Cichocki, A. (2006). Non-negative matrix factorization with quasi-newton optimization. *In International Conference on Artificial Intelligence and Soft Computing*, pages 870–879. Springer.

Titre : Estimation du niveau sonore de sources d'intérêt au sein de mixtures sonores urbaines : application au trafic routier

Mots clés : Acoustique urbaine; Séparation de sources sonores; Factorisation en matrices non-négatives; Environnement sonore urbain.

Résumé : Des réseaux de capteurs acoustiques sont actuellement mis en place dans plusieurs grandes villes afin d'obtenir une description plus fine de l'environnement sonore urbain. Un des défis à relever est celui de réussir, à partir d'enregistrements sonores, à estimer des indicateurs utiles tels que le niveau sonore du trafic routier. Cette tâche n'est en rien triviale en raison de la multitude de sources sonores qui composent cet environnement. Pour cela, la Factorisation en Matrices Non-négatives (NMF) est considérée et appliquée sur deux corpus de mixtures sonores urbaines simulés. L'intérêt de simuler de tels mélanges est la possibilité de connaitre toutes les caractéristiques de chaque classe de son dont le niveau sonore exact du trafic routier. Le premier corpus consiste en 750 scènes de 30 secondes mélangeant une composante de trafic routier dont le niveau sonore est calibré et une classe de son plus générique.

Les différents résultats ont notamment permis de proposer une nouvelle approche, appelée « NMF initialisée seuillée », qui se révèle être la plus performante. Le deuxième corpus créé permet de simuler des mixtures sonores plus représentatives des enregistrements effectués en villes, dont leur réalisme a été validé par un test perceptif. Avec une erreur moyenne d'estimation du niveau sonore inférieure à 1,2 dB, la NMF initialisée seuillée se révèle, là encore, la méthode la plus adaptée aux différents environnements sonores urbains. Ces résultats ouvrent alors la voie vers l'utilisation de cette méthode à d'autres sources sonores, celles que les voix et les sifflements d'oiseaux, qui pourront mener, à terme, à la réalisation de cartes de bruits multi-sources.

Title: Estimation of the noise level of sources of interest within urban noise mixtures: application to road traffic

Keywords: Urban acoustics; Sound source separation; Non-negative Matrix Factorization; Urban sound environment

Abstract: Acoustic sensor networks are being set up in several major cities in order to obtain a more detailed description of the urban sound environment. One challenge is to estimate useful indicators such as the road traffic noise level on the basis of sound recordings. This task is by no means trivial because of the multitude of sound sources that composed this environment. For this, Non-negative Matrix Factorization (NMF) is considered and applied on two corpuses of simulated urban sound mixtures. The interest of simulating such mixtures is the possibility of knowing all the characteristics of each sound class including the exact road traffic noise level.
The first corpus consists of 750 30-second scenes mixing a road traffic component with a calibrated sound level and a more generic sound class.

The various results have notably made it possible to propose a new approach, called 'Thresholded Initialized NMF', which is proving to be the most effective. The second corpus created makes it possible to simulate sound mixtures more representatives of recordings made in cities whose realism has been validated by a perceptual test. With an average noise level estimation error of less than 1.3 dB, the Thresholded Initialized NMF stays the most suitable method for the different urban noise environments.
These results open the way to the use of this method for other sound sources, such as birds' whistling and voices, which can eventually lead to the creation of multi-source noise maps.