Estimation du niveau sonore du trafic routier au sein de mixtures sonores urbaines par la Factorisation en Matrices Non négatives

J.-R. Gloaguen[a], A. Can[a], M. Lagrange[b] et J.-F. Petiot[b]
[a]IFSTTAR, CEREMA, UMRAE, Route de Bouaye, F-44344 Bouguenais, France
[b]Ecole Centrale de Nantes - LS2N, 1 rue de la noë, 44321 Nantes, France
jean-remy.gloaguen@ifsttar.fr

Les réseaux de capteurs acoustiques sont de plus en plus utilisés dans les villes et apparaissent comme un outil possible pour enrichir les cartes de bruit modélisées du trafic routier par des techniques d'assimilation de données ou bien pour valider des cartes modélisées par des mesures. Ces développements nécessitent tout d'abord de pouvoir isoler des mesures le niveau sonore de la circulation routière. Cette tâche est tout sauf triviale en raison des multiples sources sonores qui existent dans le milieu urbain. Dans le présent document, la Factorisation de Matrices Non-négatives est appliquée sur un corpus de scènes sonores simulé basé sur de véritables enregistrements annotés, et dont le réalisme a été validé perceptivement, en vue d'estimer les niveaux sonores du trafic routier. Les résultats démontrent l'efficacité de cette approche en estimant le niveau sonore du trafic routier avec des erreurs moyennes de moins de 1,3 dB sur l'ensemble du corpus de scènes sonores testé.

1 Introduction

La cartographie du bruit de trafic routier, principale nuisance sonore en ville, a été recommandée par la Directive européenne 2002/EC/49 afin d'estimer les niveaux sonores équivalent pondérés A, L_{DEN} (*Day-Evening-Night*) et L_N (*Night*) [1] à travers toutes les villes de plus de 100 000 habitants pour déterminer le nombre de citadins exposés à des niveaux sonores élevés [2]. Des plans d'action sont ensuite élaborés pour réduire cette exposition. Toutefois, les cartes de bruit souffrent de certaines limites dues aux simplifications générées par les outils numériques [3], par les modèles de sources et de propagation considérés ou encore par la collecte de données. De plus, les indicateurs produits masquent l'évolution des niveaux sonores due aux variations du trafic tout au long de la journée. Par conséquent, les mesures de bruit sont de plus en plus utilisées en complément des simulations pour décrire les environnements urbains bruyants. Plusieurs configurations de mesures ont été proposées au cours des dernières années, dont des mesures mobiles avec des microphones de hautes qualités [4], des mesures participatives à travers des applications smartphones dédiées [5] ou le développement de réseaux de capteurs fixes. Dans ce dernier cas, ces réseaux de capteurs peuvent être basés soit sur des capteurs de hautes qualités comme dans [6, 7], soit sur des capteurs à bas coûts comme dans le projet CENSE [8]. Tous ces protocoles de mesures permettent a priori de combiner les mesures et la modélisation pour améliorer la précision des cartes de bruit produites [9, 10]. Toutefois, ces travaux partent du principe implicite que les mesures de bruit contiennent principalement du trafic routier. Même si la circulation routière est prédominante dans certaines zones urbaines, d'autres environnements sonores urbains sont composés de nombreuses sources sonores qui ne sont pas reliées au trafic routier (voix, pas, bruits de pas, klaxons, sifflements d'oiseaux). Si ces sources ne sont pas traitées correctement, leur prise en compte entrainent de mauvaises estimation des niveaux sonores. Le recouvrement entre toutes ces sources étant fréquent, l'estimation correcte du niveau sonore du trafic dans un mélange sonore urbain reste donc difficile.

Une première approche consiste à détecter la présence des évènements sonores [11] afin de délimiter les périodes où le trafic est prépondérant [12]. Une autre approche, suivie dans cet article, consiste à considérer le paradigme de la séparation aveugle de sources, voir Figure 1. Parmi les différentes méthodes existantes (CASA, ICA), la Factorisation de Matrices Non-négatives (abrégé NMF pour *Non-negative Matrix Factorisation* en anglais) [13], semble la méthode la plus pertinente pour des capteurs monophoniques et pour traiter le problème du recouvrement des sources sonores. De nombreuses applications peuvent

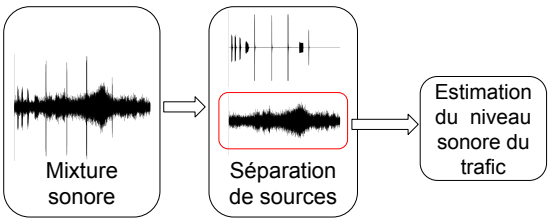

FIGURE 1 – Bloc diagramme du modèle de séparation aveugle de sources.

être trouvées pour des contenus musicaux [14] et de paroles [15]. Une première étude [16] a démontré l'intérêt de l'utilisation de la NMF, en l'appliquant sur un grand nombre de scènes sonores simulées mélangeant une composante trafic avec des sons urbains spécifiques à des niveaux sonores calibrés. L'outil doit maintenant être confronté à des scènes sonores urbaines plus réalistes.

Dans cet article, la NMF est appliquée à un corpus de scènes sonores simulées, généré à partir d'enregistrements urbains annotés, dont le réalisme a été validé par un test perceptif. La NMF et ses différentes versions implémentées sont décrites dans la section 2. Puis, le corpus des scènes sonores urbaines est présenté dans la section 3. Enfin, dans la section 4 et 5, le protocole expérimental et les résultats sont exposés.

2 Factorisation en Matrices Non-négatives

La Factorisation en Matrices Non-négatives (abrégé NMF pour *Non-negative Matrix Factorization* en anglais) [13] est une méthode d'approximation linéaire qui consiste à approximer une matrice non négative $\mathbf{V} \in \mathbf{R}^{+}_{F\times N}$ par le produit de deux matrices elles-mêmes non négatives, $\mathbf{W}$, appelée *dictionnaire*, et $\mathbf{H}$, appelée *matrice d'activation*.

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

Les dimensions de $\mathbf{W}$ et $\mathbf{H}$, respectivement $F\times K$ et $K\times N$, sont le plus souvent choisies afin que $F\times K + K\times N < F\times N$. La NMF est alors une approximation dite de rang faible. Cette condition n'est cependant pas obligatoire. Dans le domaine de l'audio, $\mathbf{V}$ est généralement considéré comme un spectrogramme d'amplitude obtenu par une Transformation de Fourier à Court Terme, $\mathbf{W}$ inclut alors des spectres audio et $\mathbf{H}$ traduit l'évolution temporelle de chaque spectre, voir Figure 2. En raison de la contrainte de non-négativité, seules les combinaisons additives entre les éléments de $\mathbf{W}$ sont

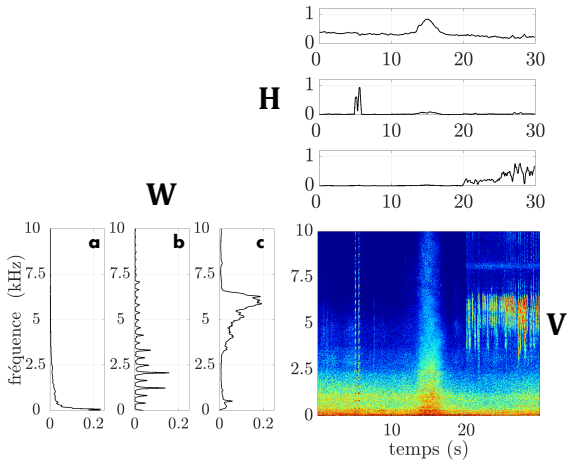

FIGURE 2 – NMF pour un échantillon audio avec 3 éléments ($\mathbf{K} = 3$) : voiture de passage (a), klaxon (b) et sifflement d'oiseau (c).

considérées menant alors à une représentation par partie. L'approximation de $\mathbf{V}$ par le produit $\mathbf{WH}$ est définie par une fonction de coût à minimiser,

$$\min_{\mathbf{H}\geq 0,\mathbf{W}\geq 0}\left[D\left(\mathbf{V}\|\mathbf{WH}\right)=\sum_{f=1}^{F}\sum_{n=1}^{N}d_{\beta}\left(\mathbf{V}_{fn}|\left[\mathbf{WH}\right]_{fn}\right)\right] \quad (2)$$

où $d_\beta(x|y)$ appartient à la classe des β-divergences, une sous-classe appartenant aux divergences de Bregman [17] qui comprend la distance euclidienne (Eq. 3a) et la divergence Kullback-Leibler (Eq. 3b),

$$d_{\beta}(x|y)=\begin{cases}\dfrac{1}{2}(x-y)^2, & \beta=2, \quad (3a)\\ x\log\dfrac{x}{y}-x+y, & \beta=1. \quad (3b)\end{cases}$$

Le problème de minimisation (2) est résolu itérativement en mettant à jour la forme des matrices $\mathbf{W}$ et $\mathbf{H}$. Les algorithmes de mises à jour multiplicatifs sont ici choisis car ils assurent des résultats non négatifs et la convergence des résultats [18].

2.1 NMF supervisée

Ici, dans un contexte urbain, les sources sonores sont connues et leurs échantillons audio peuvent être obtenus pour générer $\mathbf{W}$, voir la partie 4.1. $\mathbf{H}$ est alors la seule matrice à déterminer et est mise à jour à chaque itération (Eq. 4) [18].

$$\mathbf{H}^{(i+1)}\leftarrow\mathbf{H}^{(i)}\otimes\left(\frac{\mathbf{W}^T\left[\left(\mathbf{WH}^{(i)}\right)^{(\beta-2)}\otimes\mathbf{V}\right]}{\mathbf{W}^T\left[\mathbf{WH}^{(i)}\right]^{(\beta-1)}}\right). \quad (4)$$

Le produit $A\otimes B$ et A/B symbolisent le produit et le ratio de Hadamard. Cette première approche correspond à la NMF supervisée (NMF-SUP). Comme la position de chaque élément est indexée, la séparation de la source *trafic* des autres sources sonores est faite en extrayant, du dictionnaire et de la matrice d'activation, les éléments associés :

$$\tilde{\mathbf{V}}_{trafic}=\left[\mathbf{WH}\right]_{trafic}. \quad (5)$$

2.2 NMF semi-supervisée

Une seconde approche est considérée au travers de la NMF semi-supervisée (NMF-SEM) [15, 19] pour mieux prendre en compte les autres sources sonores. Cette méthode propose de décomposer $\mathbf{W}_{F\times(K+J)}$ tel que $\mathbf{W}=[\mathbf{W_s}\ \mathbf{W_r}]$ où $\mathbf{W_s}_{F\times K}$ est une partie fixe de $\mathbf{W}$ composée de spectres audio *trafic* et $\mathbf{W_r}_{F\times J}$ une partie mobile mise à jour, voir Eq. 7a. Il est donc possible d'y inclure des éléments non présents dans $\mathbf{W_s}$. La dimension de $\mathbf{W_r}$ est choisie telle que $J << K$ afin de considérer au mieux la source sonore présente dans $\mathbf{W_s}$. $\mathbf{H}$ est ensuite décomposée en deux matrices, $\mathbf{H}_{(K+J)\times N}=\begin{bmatrix}\mathbf{H_s}\\ \mathbf{H_r}\end{bmatrix}$. L'Eq. 1 devient

$$\mathbf{V}\approx\mathbf{WH}=\mathbf{W_sH_s}+\mathbf{W_rH_r}. \quad (6)$$

Les paramètres $\mathbf{H_r}$ et $\mathbf{H_s}$ sont mis à jour séparément, voir Eq. 7b et 7c,

$$\mathbf{W_r}^{(i+1)}\leftarrow\mathbf{W_r}^{(i)}\otimes\left(\frac{\left[\left(\mathbf{W_rH_r}^{(i)}\right)^{(\beta-2)}\otimes\mathbf{V}\right]\mathbf{H_r}^T}{\left(\mathbf{W_rH_r}^{(i)}\right)^{(\beta-1)}\mathbf{H_r}^T}\right), \quad (7a)$$

$$\mathbf{H_r}^{(i+1)}\leftarrow\mathbf{H_r}^{(i)}\otimes\left(\frac{\mathbf{W_r}^T\left[\left(\mathbf{W_rH_r}^{(i)}\right)^{(\beta-2)}\otimes\mathbf{V}\right]}{\mathbf{W_r}^T\left(\mathbf{W_rH_r}^{(i)}\right)^{(\beta-1)}}\right), \quad (7b)$$

$$\mathbf{H_s}^{(i+1)}\leftarrow\mathbf{H_s}^{(i)}\otimes\left(\frac{\mathbf{W_s}^T\left[\left(\mathbf{W_sH_s}^{(i)}\right)^{(\beta-2)}\otimes\mathbf{V}\right]}{\mathbf{W_s}^T\left(\mathbf{W_sH_s}^{(i)}\right)^{(\beta-1)}}\right). \quad (7c)$$

Cette approche a l'avantage, avec l'ajout de $\mathbf{W_r}$, d'apporter plus de flexibilité et ainsi d'être plus adaptable aux différents environnements sonores urbains. L'estimation du signal *trafic* est ensuite définie à partir de la partie fixe $\mathbf{W_s}$,

$$\tilde{\mathbf{V}}_{trafic}=\left[\mathbf{W_sH_s}\right]. \quad (8)$$

2.3 NMF initialisée seuillée

Une troisième approche est enfin proposé développée pour cette étude : la NMF initialisée seuillée (abrégé NMF-IS). Comme la source sonore cible est connue, un dictionnaire initial, $\mathbf{W_0}$, peut être conçu. Mais à l'inverse de la NMF-SUP, le dictionnaire est ensuite mis à jour alternativement avec $\mathbf{H}$,

$$\mathbf{W}^{(i+1)}\leftarrow\mathbf{W}^{(i)}\otimes\left(\frac{\left[\left(\mathbf{W}^{(i)}\mathbf{H}\right)^{(\beta-2)}\otimes\mathbf{V}\right]\mathbf{H}^T}{\left[\mathbf{W}^{(i)}\mathbf{H}\right]^{(\beta-1)}\mathbf{H}^T}\right). \quad (9)$$

Avec cette opération, le dictionnaire est mis à jour en étant initialement orienté vers la source sonore ciblée (le trafic routier) tout en pouvant être adapté spécifiquement au contenu de la scène grâce aux mises à jour. Après I itérations, chaque élément k du dictionnaire final, $\mathbf{W'}$, est comparé à sa valeur initiale dans $\mathbf{W_0}$ par un calcul de similarité cosinus $D_\theta\left(\mathbf{W_0}\|\mathbf{W'}\right)$ afin d'identifier quel élément est resté proche de la composante trafic.

$$D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right)=\frac{\mathbf{w_0}\otimes\mathbf{w'}}{\|\mathbf{w_0}\|\otimes\|\mathbf{w'}\|} \quad (10)$$

où $\mathbf{w}$ est un élément k de $\mathbf{W}$ de dimensions $F\times 1$. Lorsque $D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right)=1$, l'élément $\mathbf{w'}$ est exactement similaire à son spectre intial $\mathbf{w_0}$. Si $D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right)=0$, l'élément est complètement différent. L'extraction des éléments *trafic*

de $\mathbf{W}'$ est ensuite effectuée à l'aide d'une méthode de seuillage dur [20]. Les éléments *trafic*, $\mathbf{W}_{trafic}$, sont estimés en pondérant $\mathbf{w}'$ par α_k tel que $\mathbf{w}_{trafic} = \alpha_k \mathbf{w}'$ avec $\alpha_k = 1$ si $D_\theta(\mathbf{w}_0 || \mathbf{w}') > t_h$, sinon $\alpha_k = 0$.

Ces 3 méthodes sont appliquées sur des scènes sonores simulées afin de comparer les niveaux sonores estimés avec les solutions exactes. Pour cela, un corpus sonore réaliste est généré à partir d'enregistrements urbains mêlant de nombreuses sources sonores.

3 Création du corpus de scènes sonores réalistes

Les scènes sonores sont extraites de 74 enregistrements de 2 à 5 minutes, réalisés dans le 13 arrondissement de Paris (France) [1] dans des environnements sonores différents représentatifs du milieu urbain. Une description complète du protocole expérimental d'enregistrement se trouve dans [21]. Les enregistrements sont classés selon quatre environnements sonores différents [22] : parc (*Pa*, 8 fichiers audio), rue calme (*Ca*, 35 fichiers audio), rue bruyante (*Br*, 23 fichiers audio) et rue très bruyante (*T-Br*, 8 fichiers audio). Chaque audio est ensuite annoté afin de relever les différents évènements sonores (temps d'apparition et de disparition) et les classe de son auxquels ils appartiennent. Cette phase d'annotation sert ensuite à retranscrire les enregistrements en scènes sonores simulées qui auront ainsi la même distribution d'événements sonores que des scènes réalistes.

Les scènes sonores sont générées avec le logiciel de simulation *SimScene* [2] [23] qui génère des mixtures sonores monaurales au format wav à une fréquence d'échantillonnage de 44,1 kHz à partir d'une base de données de sons isolés. Le contrôle des paramètres de haut niveau peut être géré par l'utilisateur comme la présence d'une classe de son, le temps entre chaque échantillon d'une même classe de son, le rapport entre le niveau sonore d'une classe d'événement avec le bruit de fond... Il permet aussi la conception de scènes sonores à partir des fichiers textes d'annotations. En sortie, *SimScene* génère un fichier audio de la mixture sonore globale et un audio pour chaque classe de son présente dans la scène, ce qui permet de connaître sa contribution exacte.

Pour transcrire les enregistrements dans des scènes simulées, une base de données de sons isolés de haute qualité (format wav, fréquence d'échantillonnage de 44,1 kHz, *Rapport Signal à Bruit* élevé) a été constituée à partir d'échantillons audio trouvés en ligne (*freesound. org*) ou à l'aide d'une base de données déjà existante [24]. La base de sons est composée de deux catégories de sons : la catégorie *événement*, qui comprend 245 échantillons sonores brefs considérés comme prédominant, d'une durée de 1 à 20 secondes et classés parmi 21 classes sonores (*sifflement d'oiseau, klaxon de voiture, passage de voiture, voix, sirène* ...) et la catégorie *bruit de fond* (ou *texture*) regroupant 154 sons de longue durée (≈ 1m30), dont les propriétés acoustiques ne varient pas dans le temps. Cette catégorie comprend, entre autres, comme classes de sons : *sifflements d'oiseaux, le bruit de foule, pluie, bruit constant de la circulation* .... Chaque classe sonore est présente dans de multiples échantillons afin d'apporter de la diversité. Comme le trafic routier est la principale composante de l'environnement urbain et la source sonore d'intérêt, 103 passages de voiture ont été enregistrés sur une piste d'essai de l'Ifsttar. Les enregistrements ont été réalisés pour 4 voitures (Renault Senic, Renault Megane et Renault Clio, Dacia Sandero) pour différentes vitesses et différents rapports de vitesses. Les échantillons audio des deux premières voitures (Renault Senic et Renault Megane) sont inclus dans la base de sons de *SimScene* (50 fichiers audio au total). Les 53 extraits audio restants sont dédiés à la conception du dictionnaire dans le cadre de la NMF (voir la partie 4.1). Avec cette base de données constituée, le logiciel *SimScene* et les annotations des enregistrements, 74 scènes sonores simulées sont générées ayant la même structure temporelle que les enregistrements.

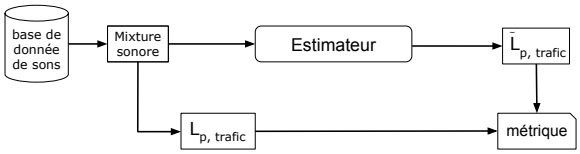

FIGURE 3 – Bloc diagramme du protocole expérimental.

Afin de valider la conception et le réalisme de ces scènes, un test perceptif a été mis en place où des auditeurs ont eu à évaluer le réalisme des scènes retranscrites au regard des enregistrement. Celui-ci révèle que les auditeurs perçoivent les scènes simulées de la même manière que les enregistrements. Une description complète de la base de données, du test perceptif et des résultats se trouvent dans [25]. Comme le corpus sonore simulé est suffisamment proche des enregistrements, ces scènes sonores peuvent être utilisées pour évaluer les performances de la NMF afin d'estimer le niveau sonore du trafic.

4 Protocole expérimental

Le protocole expérimental consiste à estimer, sur les 74 scènes disponibles classées selon les 4 environnements sonores, le niveau sonore équivalent du trafic de l'ensemble des scènes, $\tilde{L}_{p,trafic}$ (dB) et à les comparer à leurs valeurs exactes données par le processus de simulation, $L_{p,trafic}$, voir Figure 3.

Comme le trafic routier est principalement composé d'un contenu situé en basses fréquences, un premier estimateur est considéré à travers un filtre passe-bas (filtre PB). Il consiste à filtrer les scènes sonores à différentes fréquences de coupure $f_c \in \{500, 1, 2k, 5k, 10k, 10k, 20k\}$ Hz. L'énergie restante située dans la bande passante est assimilée au trafic routier,

$$\tilde{\mathbf{V}}_{trafic} = \mathbf{V}_{f_c}. \tag{11}$$

Le deuxième estimateur est basé sur les 3 formulations de la NMF présentée dans la partie 2. Entre la construction du dictionnaire et le calcul métrique, plusieurs facteurs expérimentaux, aux multiples modalités, interviennent (voir Tableau 1).

1. Ces enregistrements ont été réalisés dans le cadre du projet Grafic financé par l'Ademe

2. projet open-source disponible à l'adresse : https://bitbucket.org/mlagrange/simscene

Tableau 1 – Résumé des facteurs expérimentaux et de leurs modalités pris en compte pour estimer le niveau sonore du trafic routier.

facteurs expérimentaux	**modalités**						**nombre de modalités**
environnement sonore	parc (*Pa*)		rue calme (*Ca*)	rue bruyante (*Br*)		rue très bruyante (*T-Br*)	4
méthode	filtre PB		NMF-SUP	NMF-SEM		NMF-IS	4
$\mathbf{f_c}$ (kHz)	0.5	1	2	5	10	20	6
$\mathbf{w_t}$ (s)		0,5			1		2
K	25		50	100		200	4
β		1			2		2
valeur seuil $\mathbf{t_h}$	de 0,30 à 0,60 avec un pas d'incrément de 0,01						31

4.1 Construction du dictionnaire

Le dictionnaire **W** est conçu à partir d'une deuxième base de sons spécialement dédiée à cette tâche pour éviter tout problème de surapprentissage. Il contient les 53 fichiers audio des 2 autres voitures (Dacia Sandero et Renault Clio) enregistrés (voir partie 3).

Le spectrogramme de chaque fichier audio est d'abord calculé (fenêtre $w = 2^{12}$ avec 50 % de recouvrement). Ce spectrogramme est ensuite découpé en plusieurs trames temporelles de durée $\mathbf{w_t} = \{0.5,1\}$ seconde. Dans chaque spectrogramme découpé, la valeur *rms* sur chaque trame fréquentielle est calculée pour obtenir un spectre de dimensions $F \times 1$. Cette méthode permet de décrire l'échantillon audio avec des spectres de finesses différentes et d'obtenir différentes formes caractéristiques de spectres de trafic. À partir des 53 fichiers audio, pour $\mathbf{w_t} \in \{0,5,$ 1} secondes, 2218 et 1109 éléments sont respectivement générés. Un algorithme de clustering *K*-mean est appliqué pour réduire ces dimensions à $\mathbf{K} \in \{25, 50, 100, 200\}$ afin d'éviter des informations redondantes et de diminuer le temps de calcul. Les **K** clusters obtenus composent alors les éléments de **W**. Chaque base de **W** est ensuite normalisée telle que $\|\mathbf{w}_k\| = 1$ où $\| \bullet \|$ est la norme ℓ_1. Le Tableau 1 résume les différentes modalités des deux facteurs expérimentaux (**K** et $\mathbf{w_t}$).

4.2 Facteurs expérimentaux de la NMF

La NMF est exécutée pour 2 β-divergences : $\beta = 2$ (distance euclidienne) et $\beta = 1$ (divergence de Kullback-Leibler). Le spectrogramme **V** et le dictionnaire **W** sont exprimé en bandes de tiers d'octave qui réduisent la prédominance des hautes fréquences, où la composante trafic est absente, par son échelle logarithmique. De plus, comme le nombre de trame fréquentielle est plus faible ($F = 29$), le cout de calcul est réduit. Pour la NMF-SEM, le nombre d'éléments de $\mathbf{W_r}$ est fixé à $J = 2$. La valeur seuil, $\mathbf{t_h}$, est définie entre 0,30 et 0,60 avec un pas d'incrément de 0,01. Le résumé des facteurs expérimentaux et de leurs différentes modalités est présenté dans le Tableau 1. Le nombre de scènes disponibles par environnement sonore est défini sous la variable M ($M = 8$ pour *parc*, $M = 35$ pour *rue calme*, $M = 23$ pour *rue bruyante*, $M = 8$ pour *rue très bruyante*). Les spectres approximatifs et stabilisés du trafic $\tilde{\mathbf{V}}_{trafic}$ sont obtenus après 400 itérations. Les niveaux sonores estimés du trafic en dB, $\tilde{L}_{p,trafic}$, pour chaque association unique de modalité et pour les M scènes associées, sont alors déduits :

$$\tilde{L}_{p,trafic} = 20 \log_{10} \frac{p_{rms}}{p_0}, \tag{12}$$

avec la pression acoustique de référence, $p_0 = 2 \times 10^{-5}$ Pa.

4.3 Métrique

Les niveaux sonores du trafic, $\tilde{L}_{p,trafic}$, sont comparés aux valeurs exactes, $L_{p,trafic}$, à travers l'erreur absolue moyenne (abrégé *MAE* pour *Mean Absolute Error* en anglais). L'erreur *MAE* exprime la qualité de la reconstruction à long terme du signal et équivaut à la moyenne de la différence absolue entre le niveau sonore exact et le niveau sonore estimé,

$$MAE = \frac{\sum_{i=1}^{M} |L^i_{p,trafic} - \tilde{L}^i_{p,trafic}|}{M}. \tag{13}$$

Il est alors possible d'exprimer l'erreur *MAE* pour chaque paramètre unique mais aussi de faire la moyenne de cette métrique sur l'ensemble des 4 environnements sonores pour pouvoir estimer le réglage optimal qui offre la plus faible erreur :

$$mMAE = \frac{\sum_{i=1}^{4} MAE_i}{4}, \tag{14}$$

où les autres facteurs expérimentaux (méthode, f_c, **K**, $\mathbf{w_t}$, β, valeur seuil $\mathbf{t_h}$) sont fixes.

5 Résultats

Le Tableau 2 résume les plus faibles erreurs *mMAE* selon les paramètres **méthode** (filtre PB, NMF-SUP, NMF-SEM et NMF-IS) et β avec les autres modalités correspondantes.

L'erreur émise par le filtre pour $f_c = 20$ kHz correspond à l'erreur qui serait produite si aucun traitement n'est réalisé sur les enregistrements et que toutes les sources sonores sont prises en compte sans distinction. Cette erreur est ainsi élevée avec un écart type important. L'erreur minimale pour la méthode de référence du filtre PB est obtenue avec $f_c = 500$ Hz ($mMAE = 2{,}14$ ($\pm$ 1,83) dB). Lorsque l'on considère toutes les scènes sonores, la NMF-SUP ne permet pas d'obtenir une meilleure reconstruction que le filtre PB de 500 Hz, et cela pour toutes les valeurs β. En ajoutant la partie mobile $\mathbf{W_r}$ dans le dictionnaire, la NMF-SEM,

Tableau 2 – Erreurs *mMAE* les plus faibles selon les facteurs expérimentaux β and **méthode** (en lettre gras, la combinaison de modalités la plus performante).

méthode	f_c (kHz)	β	**K**	$\mathbf{w_t}$ (s)	$\mathbf{t_h}$	*mMAE* (dB)
filtre PB	20	-	-	-	-	3,76 (± 4,35)
filtre PB	0,5	-	-	-	-	2,14 (± 1,83)
NMF-SUP	-	1	200	0,5	-	2,79 (± 3,38)
NMF-SUP	-	2	25	1	-	2,32 (± 2,80)
NMF-SEM	-	1	200	1	-	1,94 (± 0,38)
NMF-SEM	-	2	200	1	-	2,39 (± 1,23)
NMF-IS	-	1	100	1	0,34	1,38 (± 0,88)
NMF-IS	-	**2**	**200**	**0.5**	**0,32**	**1,24 (± 1,24)**

pour $\beta = 1$, réussit à obtenir une erreur plus faible que le filtre PB à 500 Hz avec un écart type réduit (*mMAE* = 1,94 (± 0,38) dB). La NMF-IS est l'approche dont l'erreur globale est la plus faible (< 1,30 dB). Le meilleur résultat est obtenu pour la NMF-IS avec $\beta = 2$, $\mathbf{K} = 200$, $\mathbf{w_t} = 0{,}5$ s et comme valeur seuil $\mathbf{t_h} = 0{,}32$ (*MAE* = 1,24 (± 1,24) dB). Cette combinaison de modalités offre la méthode la plus adaptée à tous les environnements sonores du corpus. À partir de ces résultats globaux, les erreurs *MAE* pour les combinaisons les plus performantes selon la **méthode** (filtre LP, NMF-SUP, NMF-SEM, NMF-IS) sont comparées pour les 4 environnements sonores, voir Figure 4.

À l'exception de la NMF-SEM, toutes les méthodes montrent la même évolution d'erreur : une diminution de l'erreur avec l'augmentation de la prédominance du trafic. La NMF-SEM présente une erreur plus constante pour les 4 environnements sonores. L'erreur du filtre PB est principalement importante pour les environnements où le trafic est moins présent. Comme cette approche considère l'énergie restante comme la composante trafic, aucune distinction ne peut être faite dans la bande passante entre les différentes sources sonores non liées au trafic. À l'inverse, pour les environnements bruyants et très bruyants, les performances du filtre PB sont bonnes (*MAE* < 1 dB). Les erreurs sont alors dues à la forte suppression de l'énergie du trafic par le filtre alors qu'il devient la source sonore principale.

Malgré un dictionnaire fixe composé de spectres de trafic, la NMF-SUP ne parvient pas à identifier correctement la composante *trafic*, en particulier pour les environnements *park* (*MAE* = 6,42 dB). Avec cette méthode, comme la NMF minimise la fonction de coût (Eq. 2), les éléments *trafic* du dictionnaire sont susceptibles d'être utilisés pour modéliser les autres sources sonores. Leur utilisation à mauvais escient entraine alors de fortes erreurs. À l'inverse, pour les environnements *bruit* et *très bruyant*, la NMF-SUP identifie correctement la composante du trafic (*MAE* < 0,6 dB), source sonore principale. Dans le cas de la NMF-SEM, l'ajout du dictionnaire mobile, $\mathbf{W_r}$, permet d'inclure les autres sources sonores non présentes dans le dictionnaire. Si ce comportement est avantageux pour l'environnement *park* (*MAE* = 2,03 dB) où se trouvent beaucoup de sources différentes, il est moins avantageux pour le reste des environnements où le trafic devient prédominant. En effet, $\mathbf{W_r}$ étant non contraint, des composantes *trafic* sont susceptibles de s'y retrouver pénalisant l'estimation du niveau sonore du trafic. Finalement, la NMF-IS présente les

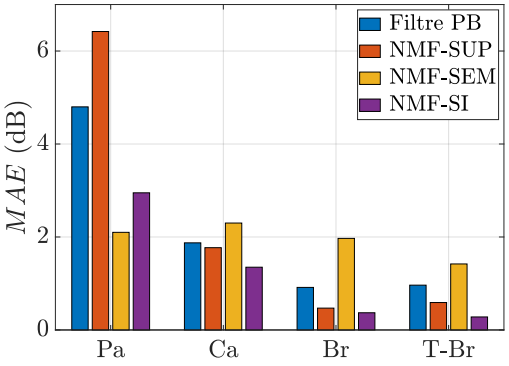

Figure 4 – Erreurs *MAE* selon chaque environnement sonore pour la meilleure combinaison du filtre PB (f_c = 500 Hz), NMF-SUP ($\beta = 2$, $\mathbf{K} = 25$, $\mathbf{w_t} = 1$ seconde), NMF-SEM ($\beta = 1$, $\mathbf{K} = 200$, $\mathbf{w_t} = 1$ seconde) et la NMF-IS ($\beta = 2$, $\mathbf{K} = 200$, $\mathbf{w_t} = 0{,}5$ seconde, $\mathbf{t_h} = 0,32$).

résultats les plus performants. L'environnement *parc* est le seul cas où une autre NMF s'est révélée significativement plus performante que la NMF-IS (*MAE* = 2,95 dB). Dans cet environnement sonore, le dictionnaire du trafic est alors composé, en moyenne, de 136 éléments. L'erreur est alors généré par la prise en compte d'éléments dont les spectres sont trop éloignés de spectres *trafic*. Pour les autres environnements sonores, la NMF-IS présente les plus faibles erreurs. Dans un environnement très bruyant, celle-ci est même très faible (*MAE* = 0,28 dB). Dans ce cas, en moyenne, $\mathbf{W}'$ est composé de 198 éléments *trafic*, soit la quasi intégralité du dictionnaire. Cette méthode surpasse donc la NMF-SUP puisque, comme le dictionnaire est mis à jour, $\mathbf{W}'$ est mieux adapté à la scène sonore que le dictionnaire fixe. L'avantage d'avoir un dictionnaire unique adapté à chaque scène sonore rend la NMF-IS très performante lorsque le trafic est prédominant. Dans le cas où le trafic est plus silencieux, l'utilisation du seuillage permet alors de ne garder que les éléments les plus pertinents.

6 Conclusion

La méthode de la Factorisation en Matrices Non-négatives a été proposée afin d'estimer le niveau sonore du trafic d'un corpus de scènes sonores urbaines simulées dont le réalisme a été vérifié par à un test perceptif.

Si la méthode semble adaptée pour de tels environnements sonores (prise en compte du recouvrement entre les multiples sources sonores présentes dans les villes, compatibilité avec des réseaux de capteurs monophoniques), l'approche supervisée et semi-supervisée ne permettent pas de reconstruire la composante *trafic* de manière suffisante. La NMF supervisée, avec son dictionnaire fixe, se révèle trop rigide pour être efficace quand le trafic est peu présent, alors que l'approche semi-supervisée avec la présence d'une partie mobile dans le dictionnaire se révèle efficace pour les environnements *park* mais échoue sur les scènes bruyantes car non-contrainte. L'approche proposée, à savoir la NMF initialisé seuillée, est finalement la méthode la plus efficace sur l'ensemble des environnements sonores avec $\beta = 2$, $\mathbf{K} = 200$, $\mathbf{w_t} = 0{,}5$ seconde et $\mathbf{t_h} = 0,32$. D'un dictionnaire composé de spectres de trafic, mis à jours, elle

permet de généraliser un dictionnaire initiale à chaque scène sonore. D'autres analyses sont nécessaires pour étendre la méthode proposée à d'autres sources sonores, comme les oiseaux ou les sons de voix, en remplaçant ou en ajoutant des éléments dans le dictionnaire construit. Cette utilisation s'avérerait utile dans le contexte de la cartographie multi-sources du bruit qui gagne en intérêt [26]. D'autres analyses sur différents corpus de scènes sonores sont nécessaires pour tester la robustesse de la méthode et sélectionner les approches les plus pertinentes pour des environnements sonores spécifiques (prédominance des sons de l'eau ou industriels, environnements ruraux ...).

À des fins de reproductibilité, le protocole expérimental, les programmes développés sous le logiciel Matlab [3] et le corpus de scènes sonores urbaines réalistes [4] sont disponibles en ligne.

Remerciements

Les auteurs souhaitent remercier Catherine Lavandier et Pierre Aumond de l'Université de Cergy-Pontoise pour avoir accepté de nous transmettre les enregistrements du projet Grafic.

Financement

Ces travaux ont été possibles avec le financement de la région Pays de la Loire.

Références

[1] S. Kephalopoulos, M. Paviotti, and F. A. Ledee. Common noise assessment methods in europe (cnossos-eu), 2012.

[2] C. Nugent, N. Blanes, J. Fons, et al. Noise in europe 2014. *European Enviroment Agency*, 10 :2014, 2014.

[3] H. Van Leeuwen and S. Van Banda. Noise mapping-state of the art-is it just as simple as it looks ? *Proceedings of EuroNoise 2015*, 2015.

[4] A. Can, L. Dekoninck, and D. Botteldooren. Measurement network for urban noise assessment : Comparison of mobile measurements and spatial interpolation approaches. *Applied Acoustics*, 83 :32–39, 2014.

[5] J. Picaut, P. Aumond, A. Can, et al. Noise mapping based on participative measurements with a smartphone. In *Acoustics '17 Boston*, volume 141 of *The Journal of the Acoustical Society of America*, page 3808, Boston, United States, June 2017.

[6] C. Mietlicki, F. Mietlicki, and M. Sineau. An innovative approach for long-term environmental noise measurement : Rumeur network. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2012, pages 7119–7130. Institute of Noise Control Engineering, 2012.

[7] P. Maijala, Z. Shuyang, T. Heittola, and T. Virtanen. Environmental noise monitoring using source classification in sensors. *Applied Acoustics*, 129 :258–267, 2018.

[8] J. Picaut, A. Can, J. Ardouin, et al. Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling. In *173rd Meeting of the Acoustical Society of America and the 8th Forum Acusticum (Acoustics'17)*, pages 3808–3808, Boston, MA, United States, June 2017.

[9] R. Ventura, V. Mallet, V. Issarny, et al. 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, 2017.

3. https://github.com/jean-remyGloaguen/articleNmfTrafficSimScene2018

4. https://zenodo.org/record/1184443#.WqfVb3wiGpo

[10] W. Wei, T. Van Renterghem, B. De Coensel, and D. Botteldooren. Dynamic noise mapping : A map-based interpolation between noise measurements with high temporal resolution. *Applied Acoustics*, Complete(101) :127–140, 2016.

[11] G. Parascandolo, H. Huttunen, and T. Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444, March 2016.

[12] J. C. Socoró, F. Alías, and R. M. Alsina-Pagès. An Anomalous Noise Events Detector for Dynamic Road Traffic Noise Mapping in Real-Life Urban and Suburban Environments. *Sensors*, 17(10) :2323, October 2017.

[13] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791, October 1999.

[14] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180, October 2003.

[15] G. J. Mysore and P. Smaragdis. 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, 2011.

[16] J-R. Gloaguen, M. Lagrange, A. Can, and J-F. Petiot. Estimation of road traffic sound levels in urban areas based on non-negative matrix factorization techniques, submitted for publication. 2018.

[17] R. Hennequin, B. David, and R. Badeau. Beta-Divergence as a Subclass of Bregman Divergence. *IEEE Signal Processing Letters*, 18(2) :83–86, February 2011.

[18] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9) :2421–2456, 2011.

[19] H. Lee, J. Yoo, and S. Choi. Semi-Supervised Nonnegative Matrix Factorization. *IEEE Signal Processing Letters*, 17(1) :4–7, January 2010.

[20] D. L. Donoho and I. M. Johnstone. Threshold selection for wavelet shrinkage of noisy data. In *Engineering in Medicine and Biology Society. Engineering Advances : New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, volume 1, pages A24–A25. IEEE, 1994.

[21] P. Aumond, A. Can, B. De Coensel, D Botteldooren, C. Ribeiro, and C. Lavandier. Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context. *Acta Acustica united with Acustica*, 103(1–1), 2016.

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

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

[24] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044. ACM, 2014.

[25] J-R. Gloaguen, A. Can, M. Lagrange, and J-F. Petiot. Creation of a corpus of realistic urban sound scenes with controlled acoustic properties. In *173rd Meeting of the Acoustical Society of America and the 8th Forum Acusticum (Acoustics'17)*, pages 4044–4044, Boston, MA, United States, June 2017.

[26] P. Aumond, L. Jacquesson, and A. Can. Probabilistic modeling framework for multisource sound mapping, submitted for publication. 2017.