Reconnaissance et estimation du bruit de trafic dans l'environnement urbain

Jean-Rémy Gloaguen,
Arnaud Can
LUNAM Université
IFSTTAR Centre de Nantes
AME-LAE
Route de Bouaye
CS 4
44344 Bouguenais Cedex
France
E-mail :
jean-remy.gloaguen@ifsttar.fr
arnaud.can@ifsttar.fr

Mathieu Lagrange,
Jean-François Petiot
LS2N équipe SIMS UMR CNRS 6597
Ecole Centrale de Nantes
1 rue de la Noë
44321 Nantes
France

Résumé
La reconnaissance des sources sonores est un élément clef de la « Smart City » au service de l'environnement sonore. Elle vient en soutien aux réseaux de capteurs qui se déploient dans de nombreuses villes européennes, en améliorant la connaissance qualitative des sources de bruit qui composent les environnements sonores. Les méthodes de reconnaissance des sources étudiées permettront à terme de faire le lien entre cartes de bruit mesurées et modélisées, en estimant de manière automatique la contribution, par exemple, du trafic routier dans des mesures réalisées. Elles rendent également possible une cartographie des environnements sonores par type de source d'intérêt. À ce titre, cet article présente l'application de la factorisation en matrices non-négative (NMF) à la reconnaissance du bruit de trafic routier dans des mixtures sonores urbaines. Cette méthode est particulièrement adaptée car elle n'est pas mise en défaut dans le cas très fréquent où plusieurs sources se superposent. La méthode est testée sur un corpus de scènes sonores simulées, de manière à respecter les caractéristiques acoustiques de scènes réelles tout en disposant de mesures étalons permettant de valider les choix de conception de l'estimateur. Le réalisme des scènes produites est validé perceptivement. Les résultats de la méthode sont ensuite commentés.

Abstract
Sound recognition is a key element of smart cities dedicated to sound environment, as it helps understanding from which sound sources the urban sound mixtures recorded by sensor networks are composed. By determining automatically the contribution of road traffic within the measures issued from sensors, we believe that in the near future, sound recognition will link sound maps produced 1) by models and 2) by measures. Sound recognition will also enable the designs of sound maps of a new kind, one per sound source of interest. This paper presents the adaptation of the non-negative matrix factorization (NMF) to sound recognition of road traffic in urban sound mixtures. This method is of particular interest since it can deal with overlapping sounds, which are very common in real world settings. The method is tested on a corpus of urban sound scenes generated artificially, preserving as much as possible the acoustical properties of real sound scenes. To validate this aspect, the realism of the generated scenes is validated perceptually. The results of the method are finally commented.

La mise en place de la Directive européenne 2002/49/CE a fait de la cartographie du bruit un élément incontournable de lutte contre les nuisances sonores [1]. Les guides méthodologiques publiés au cours des dernières années permettent, en principe, une harmonisation des méthodes utilisées [2]. Les cartes produites s'articulent autour d'un recensement des sources de bruit principales (trafics routier, ferroviaire et aérien, ainsi que les principales industries), un recueil et une mise en forme des variables d'entrée (topographie, bâti, etc.), suivi d'un calcul des émissions et de la propagation du bruit conforme aux normes en vigueur [3][4]. Si les cartes du bruit permettent de communiquer efficacement avec les usagers de la ville sur les niveaux d'exposition, la critique est souvent faite des limites liées à la modélisation (approximations inévitables pour un calcul sur un domaine aussi grand qu'une agglomération) et aux données d'entrée (volumes de trafic, hauteur des bâtiments, données de végétation, etc.) qui peuvent générer des approximations difficiles à quantifier. En outre, les calculs se restreignent aux sources de bruit principales, offrant une cartographie partielle des environnements sonores réels, qui sont souvent enrichis de sources sonores très variées : oiseaux, voix, travaux de voirie, cloches d'église, etc.

En parallèle, certaines villes européennes se sont dotées ces vingt dernières années de réseaux de capteurs dont la fonction est un suivi continu des environnements sonores, à l'image d'Acoucité à Lyon, ou BruitParif à Paris. Ces réseaux s'affichent comme de véritables observatoires du bruit, donnant accès pour chaque point de mesure, à l'évolution temporelle des niveaux sonores.

Ils permettent par exemple de comparer les environnements sonores à plusieurs années d'intervalle, ou d'offrir en direct un suivi des dépassements des valeurs seuils, comme le propose le site Rumeur de Bruitparif[1]. Enfin, des méthodes de cartographie du bruit basées directement sur des mesures mobiles, et bientôt peut-être sur des mesures participatives, via des applications smartphones telles que NoiseCapture[2], voient le jour. Elles permettent d'affiner la finesse spatiale des cartes et présentent l'avantage d'être sensibles à l'intégralité des sources sonores.

La mise en commun de ces deux approches permettra à terme de faire converger les cartes de bruit vers des estimations du niveau sonore plus précises en s'appuyant à la fois sur la représentation continue dans l'espace de la modélisation et sur la précision et le suivi temporel des données fournies localement par des réseaux de mesures. C'est notamment l'objet du projet CENSE (Caractérisation des environnements sonores urbains : une approche globale combinant données libres, mesures et modélisation[3]) qui propose de s'appuyer sur des méthodes d'assimilation de données[4] et sur un réseau dense de capteurs à faible coût pour faire converger les cartes de bruit de trafic simulées vers des niveaux plus réalistes. Ceci nécessite néanmoins de savoir discriminer, parmi les mesures collectées, la composante liée au bruit de trafic (modélisée dans les cartographies classiques) des autres sources de bruit (trafic aérien, activités humaines, sources naturelles…). En outre, savoir déterminer la contribution des différentes sources sonores au sein d'une mesure acoustique donnée permettrait à terme le développement d'une nouvelle génération de cartes de bruit sur lesquelles il deviendrait possible de choisir les sources sonores représentées, se rapprochant ainsi des environnements sonores perçus par les usagers de la ville. Les nombreuses applications possibles offertes par les méthodes de reconnaissance des sources expliquent l'engouement actuel autour de cette thématique de recherche. Des premiers travaux ont eu pour but de classifier les environnements sonores en fonction de leurs caractéristiques acoustiques [5][6]. Des méthodes de reconnaissance des sources s'appuyant sur des réseaux de neurones permettent quant à elles d'identifier la source de bruit prédominante au sein d'un enregistrement sonore [7]. Seulement, si ces méthodes s'avèrent efficaces pour des sons pris isolément, elles trouvent leur limite pour des environnements sonores bruités, tels que les mixtures sonores urbaines, souvent composées d'une multitude de sources superposées.

Cet article propose de générer un estimateur du niveau sonore du bruit du trafic routier à partir de la méthode de Factorisation en Matrices Non-négatives (NMF). Cette méthode, qui a déjà été appliquée avec succès dans d'autres domaines que l'acoustique environnementale, présente l'avantage de pouvoir traiter des mixtures sonores formées de nombreuses sources se superposant et d'estimer la contribution de chacune des sources présentes. La méthode et les développements spécifiques qu'elle nécessite pour être appliquée au domaine de l'acoustique environnementale, sont tout d'abord présentés. La NMF est ensuite appliquée sur un corpus de scènes sonores conçues de manière à reproduire les caractéristiques d'environnements sonores typiques (rue calme, rue bruyante, parc) et validé perceptivement. Les résultats de la NMF sur ce corpus de scènes sonores sont finalement commentés.

La Factorisation en Matrices Non-Négatives (NMF)

Principe de la NMF

La factorisation en matrice non-négative (abrégé NMF pour « *Non-negative Matrix Factorization* ») est une technique d'approximation linéaire visant à approximer une matrice **V** non-négative de dimensions F x N en un produit de deux matrices :

$$\mathbf{V} \approx \mathbf{WH} \quad , \tag{1}$$

où **W** et **H**, appelées *dictionnaire* et *activateurs*, sont deux matrices, également non-négatives, de dimensions respectives F x K et K x N. Afin d'être une méthode d'approximation dite de « faible rang », le choix de la dimension K est déterminé afin que la relation FK + KN << FN soit respectée.
Cette méthode, si elle fut introduite pour la première fois par Paatero en 1994 [8], a été popularisée à partir de 1999 à travers les travaux de Lee et Seung [9], publié dans la revue *Nature*. Présentée comme une technique d'apprentissage du dictionnaire, la NMF, par sa contrainte de non-négativité, considère la description d'éléments comme une somme purement additive de fonctions élémentaires. En d'autres termes, chaque colonne de la matrice **V** est approximée par la somme des éléments de **W** pondérés par la colonne n de la matrice **H** :

$$\mathbf{v_n} \approx \mathbf{Wh_n} \quad , \tag{2}$$

où les caractères minuscules représentent des vecteurs, là où les majuscules résument des matrices.

La NMF trouve son utilité dans de nombreuses applications dans des domaines variés (imagerie [10], en traitement de la langue [11]). Dans le domaine de l'audio, c'est Smaragdis et Brown [12] qui, les premiers, l'ont utilisée afin de retranscrire la partition d'un morceau de musique polyphonique. Le dictionnaire **W** y est alors constitué d'un ensemble de spectres résumant les notes présentes dans le morceau de musique, et les activateurs **H** résument l'évolution temporelle de chacun des spectres (exemple en Figure 1).

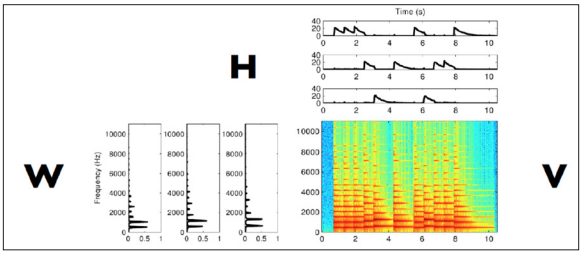

Fig. 1 : Exemple d'une NMF pour un signal audio (Au clair de la Lune) où W et H sont constitués de 3 éléments (K = 3). W représente le dictionnaire, ici un spectre pour chaque note, et H les activateurs temporels, permettant ici de reconstruire la mélodie [13]

1-http ://rumeur.bruitparif.fr/
2- http ://noise-planet.org/fr/noisecapture.html
3- http ://cense.ifsttar.fr/
4- L'assimilation de données recouvre un ensemble d'outils mathématiques permettant de corriger un champ simulé (ici la carte de bruit produite) à partir d'observation ponctuelles (ici les mesures). L'assimilation de données a été développée initialement en météorologie et en océanographie, puis plus récemment en pollution atmosphérique. Son développement dans le cadre de l'acoustique environnementale constitue un défi qui est relevé dans le cadre de CENSE.

Pour un signal audio issu d'un environnement sonore urbain, la NMF consiste à approximer son spectrogramme de puissance **V**, obtenu, par exemple, par une Transformée de Fourier à Court Terme, à l'aide du dictionnaire **W** constitué d'un ensemble de spectres provenant des différentes sources sonores présentes (trafic routier, oiseaux, klaxon…) et des activateurs **H** correspondant.

Algorithmes de mise à jour

Le problème à résoudre lors d'une NMF revient à celui d'un problème de minimisation de la distance[5] entre **V** et **WH**, avec pour objectif de faire converger le produit **WH** vers le spectrogramme **V** (équation 3)

$$\min D(\mathbf{V} \mid \mathbf{WH}) \text{ avec } \mathbf{W} \geq 0,\ \mathbf{H} \geq 0 \quad , \qquad (3)$$

Différentes formulations de cette distance existent, qui peuvent influencer la qualité de reconstruction du signal (distance euclidienne, divergence d'Itakura-Saïto, divergence de Kullback-Leibler, etc.). Chacune de ces formulations a ses caractéristiques mathématiques propres, et des algorithmes de résolution ont été proposés pour chacune d'elles, en s'appuyant sur des gradients de descente ou des heuristiques multiplicatifs. Ces algorithmes permettent alors de modifier progressivement l'allure de **W** ou **H** afin de satisfaire l'équation 3. Le lecteur trouvera plus détails dans les références suivantes : [14], [15].

Plusieurs formes de la NMF ont été proposés dans la littérature, en fonction de la connaissance totale, partielle, ou nulle, des éléments du dictionnaire. Dans le cas présent, les éléments du dictionnaire **W** sont supposés connus, et **H** seul doit être déterminé : on réalise alors une NMF dite « supervisée ». La constitution d'un dictionnaire représentatif des sources impactant l'environnement sonore est donc un enjeu crucial de la NMF « supervisée » utilisée ici. L'objectif est alors de déterminer les paramètres optimaux qui permettent de déterminer au mieux le niveau sonore du trafic routier (conception du dictionnaire, choix de la définition de la distance entre le spectrogramme estimé et mesuré, etc.). Pour cela, la NMF est appliquée, en faisant varier ces différents paramètres, sur des corpus de scènes sonores simulées où le niveau sonore du trafic est connu.

Création d'un corpus de scènes sonores réalistes

La validation des méthodes de reconnaissance des sources se heurte au fait qu'il est en pratique impossible de connaître, sur des enregistrements réels, la contribution de chacune des sources… De ce fait, les validations ne peuvent être opérées que sur des scènes artificielles. La génération de scènes sonores au contenu contrôlé, et malgré tout réalistes, constitue donc une étape cruciale pour la validation de toute méthode de reconnaissance des sources. Dans cet article, cette étape est obtenue en s'appuyant sur l'outil de génération de scènes sonores *simScene* [16]. Les scènes générées sont calquées sur des scènes réelles écoutées et annotées. Enfin, un test perceptif vient valider le réalisme des scènes générées, rendant de ce fait généralisables les résultats obtenus.

Présentation de *simScene*

Le logiciel *simScene*[6] est un simulateur de scènes sonores qui consiste à superposer des évènements sonores, issus d'une base de données de sons isolés, à un signal bruit de fond. *simScene* permet de renseigner plusieurs paramètres 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 entre chaque audio d'une même classe de son

Chaque paramètre est également complété par un écart type permettant d'instaurer de la variabilité entre les scènes simulées. Une mixture sonore créée peut également générer un audio pour chaque classe de son présent dans la scène permettant de connaitre la contribution exacte d'une classe de son présente dans la scène. Dans notre cas, ce sont toutes les classes de sons relatifs au trafic routier qui nous intéressent et qui permettent d'estimer son niveau sonore exact dans la scène.
simScene possède deux modes pour générer ces scènes. Dans le mode « *generate* », l'utilisateur renseigne lui-même les échantillons sonores présents dans la scène et chaque paramètre permettant de créer des scènes complètement artificielles (voir exemple Figure 2). À l'inverse, dans le mode « *replicate* », le schéma de la scène s'appuie sur un fichier texte où la position des évènements (début et fin) et leur classe de son correspondante sont détaillées. Ce mode permet de reproduire des scènes réelles annotées où la position de chaque évènement est connue. Afin de s'approcher au plus de scènes réalistes, c'est ce mode « *replicate* » qui a été retenu dans cette étude.

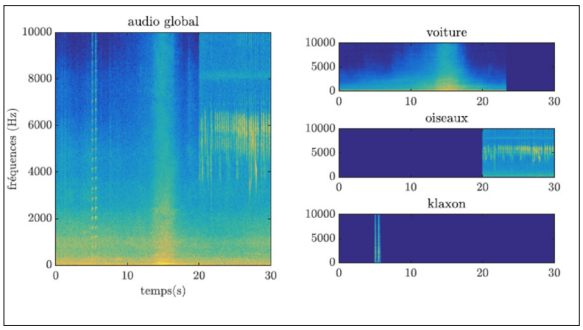

Fig. 2 : Scène créée avec simScene. À gauche, le spectrogramme du signal global composé de trois sources (une voiture, un klaxon, un oiseau), à droite les spectrogrammes des trois sources audio utilisées

Création des scènes sonores

Si *simScene* offre de nombreux paramètres pour créer de multiples scènes sonores variées, la création de scènes sonores réalistes nécessite d'une part d'avoir une base de données de sons isolés suffisamment exhaustive et représentative des environnements sonores simulés, et d'autre part de connaitre la composition des scènes sonores.

5- On parle de distance par abus de langage, le terme mathématique exact étant divergence.
6- https://bitbucket.org/mlagrange/simscene

Création d'une base de données

La base de données de sons pour *simScene* comprend un ensemble de classes de sons isolés (oiseaux, voiture, klaxon…) qui contiennent chacune plusieurs échantillons (oiseaux01.wav, oiseaux02.wav, etc.) pour permettre une grande variabilité dans les mixtures sonores créées. La plupart des échantillons ont été trouvés sur des sites en ligne de sons (freesound.org, universalsoundbank.com) et à l'aide de la base de données *UrbanSound8k* [17]. Ces éléments permettent de créer le dynamisme sonore dans les scènes et sont les évènements sonores qui sont annotés dans une scène. La base de données comprend également des sons de durées plus longues (1 min à 2 min) qui vont permettre de créer un bruit de fond sonore utile à la création de l'ambiance sonore générale (oiseaux, voix d'enfant dans une cour de récréation, trafic routier continu...).

La création de la base de données de bruit de trafic a fait l'objet d'une attention toute particulière. Des enregistrements de passages de véhicules ont été réalisés sur la piste d'essai de l'Ifsttar de Bouguenais afin de posséder un ensemble varié et maitrisé de vitesses et de modèles de véhicules. Pour chacun des quatre véhicules considérés, des mesures ont été réalisées en vitesses stabilisées à différentes vitesses et rapports de boîtes, ainsi qu'en situation de freinage et d'accélération (26 passages par véhicule). C'est finalement 104 extraits sonores qui ont pu être réalisés et qui servent à constituer la composante trafic routier des scènes sonores.

Au total, la base de données est composée de 245 échantillons d'évènements sonores provenant de 21 classes de sons (cloches, oiseaux, balais, klaxon, voiture, marteau et perceuse, toussotement, aboiement de chien, portière de voiture et de maison, avion, sirène, bruit de pas, orage, bruit de rue, roulement de valise, train et tramways, camion et voix) et de 154 échantillons de bruit de fond sonores appartenant aux classes de sons (oiseaux, bruit de chantier, foule, parc, pluie, cours de récréation, bruit de trafic continu, ventilation et vent).

Reproduction des scènes réelles

Dans un deuxième temps, il convient de générer des scènes réalistes à partir de ces échantillons. Pour cela, des enregistrements sonores réalisés dans le 13e arrondissement de Paris dans le cadre du projet GRAFIC [18] (74 enregistrements audio de 1 à 4 minutes), ont été annotés, pour déterminer les classes de sons qui y sont présentes, leurs récurrences et leur intensité. Ces écoutes ont permis de classer les enregistrements en quatre classes d'environnements distincts et représentatifs des ambiances sonores urbaines, à savoir parc, rue calme, rue animée, et rue très animée. La seconde étape a été d'établir pour les différentes ambiances sonores le niveau sonore moyen des scènes, les bruits de fonds sonores les caractérisant et les différents types de source sonores présentes ainsi que le nombre d'occurrences (nombre d'évènements / minute) de chacune des classes présentes (Tableau 1).

Les 74 enregistrements audio sont enfin reproduits par le mode replicate de *simScene* à l'aide de la première moitié des échantillons de la base de données constituée (la seconde moitié servant à constituer le dictionnaire **W** (voir partie Validation de l'approche)).

Validation perceptive des scènes

Afin de vérifier que leur rendu global est réaliste, les scènes générées sont soumises à un test perceptif [19]. Ce test consiste à faire écouter, à un panel de 50 auditeurs, un ensemble de scènes sonores comprenant autant d'enregistrements sonores réels que de scènes reconstituées. Parmi les 74 enregistrements disponibles, vingt ont été retenues composées de cinq scènes d'ambiance *parc*, six d'ambiance *rue calme*, quatre *rue animée* et cinq de *rue très animée*. Les vingt mêmes scènes reproduites sous *simScene* viennent compléter le corpus de test. La durée de chaque audio est ensuite réduite aux mêmes 30 secondes. Comme il n'est pas possible de faire écouter les 40 scènes à tous les auditeurs (la durée du test serait trop longue), chacun écoute un ensemble réduit composé de dix scènes audio réelles et dix « répliquées ».

Environnement sonore	Bruit de fond	Évènement sonore	nombre d'évènement/min
Parc	Voix sifflement des oiseaux	Voiture Voix Sifflement d'oiseaux Bruit de rue	0,5 0,5 0,5 0,5
Rue calme	Trafic routier Sifflement des oiseaux	Voiture Voix Bruit de rue Bruit de pas	1,0 0,7 0,7 0,5
Rue animée	Trafic routier	Voiture Voix Bruit de pas Bruit de rue	9,0 0,6 0,5 0,4
Rue très animée	Trafic routier	Voiture Voix Klaxon Bruit de pas	40,0 0,3 0,3 0,3

Tabl. 1 : Description des quatre classes de sons les plus récurrentes dans les environnements sonores

Un Bloc Équilibré Incomplet permet alors de définir l'ordre d'écoutes par auditeur mélangeant les scènes réelles et mélangées [20]. Enfin, pour éviter que l'auditeur ait à modifier le niveau sonore entre les scènes, toutes ont été normalisées au même niveau sonore de 65 dB.
Durant le test, l'auditeur doit alors évaluer, sur une échelle à sept points allant de « très peu réaliste » à « extrêmement réaliste », le réalisme de chaque scène écoutée. L'objectif est que l'ensemble des scènes reproduites par *simScene* ait une note moyenne similaire aux scènes réelles. Le test a été diffusé en ligne et a nécessité 12 jours pour atteindre les 50 participants. Le panel était constitué de 31 hommes et 18 femmes (1 non documenté), avec une moyenne d'âge de 36 (±12) ans.
Pour déterminer si la distribution des notes des scènes simulées est similaire aux scènes réelles, un test statistique de Student est réalisé. Celui-ci révèle, au travers de la valeur-p calculée, inférieure au seuil de signification de 5%, que les distributions sont similaires. L'estimation des notes moyennes des scènes réelles et simulées sont également très similaires ($m_{simul.}$ = 5,1 (±1,6), $m_{real.}$ = 4,9 (±1,6)). Les scènes réelles et les scènes simulées ne peuvent pas donc pas être distinguées par les auditeurs, validant ainsi la qualité perceptive du corpus de scènes sonores.

Validation de l'approche

Création du dictionnaire

Le nombre d'éléments K du dictionnaire est fixé à 100. Deux formes de dictionnaires sont comparées, selon que le dictionnaire comprend uniquement des spectres trafic ou plusieurs classes de sons (appelé *complet*). Les fichiers audio utilisés sont ceux issus de la seconde moitié de la base de données (n'ayant pas servie à la constitution des scènes), afin que le dictionnaire et les scènes sonores ne partagent pas des signaux communs. Précisons que l'intégralité des spectres des fichiers audio de la base de données pouvant excéder le nombre K, une étape de *clustering*, par l'algorithme des *k-means*, est ajoutée afin de réduire le nombre d'éléments à la dimension souhaitée, tout en s'assurant de la représentativité des éléments retenus.

Réalisation de la NMF

Parmi les scènes reproduites, huit scènes sont sélectionnées pour chaque environnement sonore (*parc, rue calme, rue animée, rue très animée*). Comme la quasi-totalité de l'énergie spectrale du trafic est inférieure à 5 kHz, les signaux sont tout d'abord filtrés avec un filtre passe-bas à 1 kHz, 2 kHz ou 5 kHz, afin de mieux se concentrer sur l'intervalle fréquentiel cible. Une méthode de base dite « filtre » consiste à simplement considérer que la contribution du trafic routier est le signal filtré. L'idée sous-jacente est que l'énergie au-delà la fréquence critique peut, de fait, être considérée comme appartenant à d'autres sources qu'au trafic routier, et que le filtre passe-bas suffit à discriminer la contribution du trafic routier du reste du signal.

La NMF est ensuite appliquée sur ces signaux filtrés pour affiner l'estimation, et mesurer l'apport de la méthode face à la méthode « filtre ». L'optimisation est réalisée en 400 itérations, pour les deux types de dictionnaire (*trafic* ou *complet*) et en fonction de la fréquence de coupure f_c du filtrage (soit 2*3 = 6 combinaisons). Les résultats sont comparés pour chaque type d'environnement sonore.

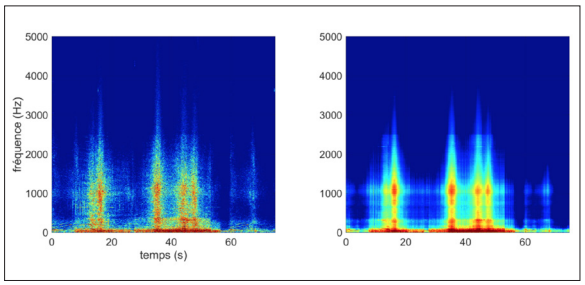

Fig. 3 : À gauche le spectrogramme V (window = 2^{14}, noverlap = 75%, nfft = 2^{14}) d'une scène appartenant à l'ambiance rue animée filtrée à 5 kHz, à droite le spectrogramme WH obtenu avec une NMF supervisée avec une distance euclidienne

Le spectrogramme du trafic est alors reconstitué en sélectionnant les spectres relatifs au trafic du dictionnaire et les activateurs correspondant,

$$\mathbf{V_{traffic}} = [\mathbf{WH}]_{\mathbf{traffic}}, \quad (4)$$

puis celui-ci est converti en un signal sonore à l'aide de la phase du signal d'origine **V** .Le niveau sonore équivalent en dB sur la totalité de la scène, $\tilde{L}_{eq}$ est ensuite calculé ainsi que les niveaux équivalents avec une résolution temporelle de 1 seconde (Figure 4)

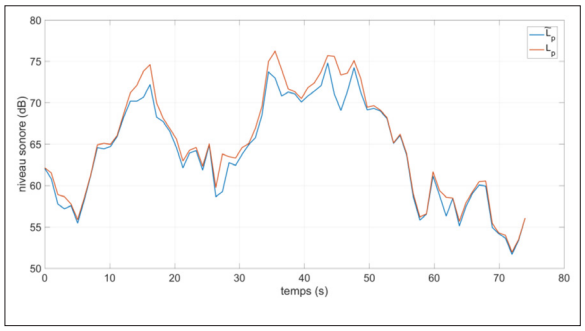

Fig. 4 : Évolution des niveaux sonores estimés, $\tilde{L}_{p,1s}$, et exacts, $L_{p,1s}$ (en dB) de la scène de la Figure 3

Pour chaque combinaison de paramètres et chaque environnement sonore, les niveaux sonores estimés $\tilde{L}_{eq}$ sont comparés aux niveaux exactes L_{eq} sur l'ensemble des N scènes par le calcul du RMSE.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(L_{eq}^{i} - \tilde{L}_{eq}^{i}\right)^2}$$

De la même manière, l'erreur RMSE est calculée entre les niveaux sonores exactes et ceux issus des signaux filtrés à la fréquence f_c.

Résultats

Dans un premier temps, l'erreur RMSE moyenne sur l'ensemble des environnements sonores est calculée (Tableau 2) La meilleure méthode pour estimer le niveau du trafic routier, sur l'ensemble des environnements sonores, est la NMF supervisée s'appuyant sur un dictionnaire composé d'éléments uniquement trafic, quelle que soit la fréquence de coupure.

RMSE			
f_c	Composition W	NMF supervisée	Filtre
1000	complet	5,02 (±1,35)	2,20 (±2,52)
	trafic	**1,77 (±1,02)**	
2000	complet	4,42 (±1,23)	2,74 (±3,04)
	trafic	**1,69 (±1,71)**	
5000	complet	4,04 (±1,09)	3,21 (±3,49)
	trafic	**1,83 (±1,92)**	

Tabl. 2 : Erreur RMSE produite sur l'ensemble de scènes pour la NMF supervisée et les signaux filtrés

A l'inverse, pour un dictionnaire composé de l'ensemble des classes de son, la NMF produit des erreurs plus fortes. Le choix d'intégrer une multitude de spectres provenant de différentes sources sonores ne permet donc pas de retrouver la contribution du trafic routier. Ceci s'explique par le fait que la NMF, minimisant la distance entre le spectrogramme du signal et le produit **WH** (équation (3)), est susceptible d'utiliser certains spectres qui n'appartiennent pas à la classe trafic, mais qui partagent des bandes fréquentielles communes, pour recomposer cette partie du signal (équation (4)). Lors de la reconstruction du signal trafic, ces éléments activés ne peuvent alors pas être pris en compte, générant une erreur plus importante.
Dans un second temps, on résume les erreurs RMSE minimales pour la méthode NMF et la méthode « filtre » pour chaque environnement sonore (Tableau 3).

Dans les environnements sonores où le trafic est la source sonore principale (*rue animée*, *rue très animée*), l'erreur des deux méthodes est faible (< 1 dB) : dans le cas des filtre passe-bas l'hypothèse que l'intégralité du signal non filtré correspond au signal *trafic* est alors vraie et pour la NMF, composée d'éléments *trafic* dans le dictionnaire, la reconstruction du signal est alors plus aisée. Mais ces environnements triviaux ne sont pas les plus intéressants en terme de reconnaissance des sources.

En revanche, dans les environnements peu touchés par le bruit de circulation, la NMF améliore significativement l'estimation des niveaux sonores trafic. Cela s'observe notamment dans les environnements sonores parc où la présence du trafic est faible par rapport aux autres sources sonores (voix, oiseaux, bruit divers…) : le recours à la NMF permet alors une diminution du RMSE de 5,9 dB à 3,3 dB. Car là où le filtre passe-bas considère que l'intégralité de la partie non filtrée correspond au signal trafic, la NMF considère le signal dans son intégralité et recherche à chaque trame temporelle la présence éventuelle des éléments « trafic » du dictionnaire.

Discussion

L'analyse montre que l'apport de la NMF est conséquent pour les scènes où les environnements sonores sont variés et où le trafic routier n'est pas la source prépondérante. En ce sens, la méthode répond bien à l'objectif initial, puisque ces environnements sonores sont ceux pour lesquels l'écart entre mesure et modélisation devient problématique lors du recalage des cartes de bruit par la mesure.

Des pistes de recherche sont en cours pour améliorer l'estimation de la contribution du trafic routier par l'application de différentes variantes de la NMF, à savoir l'introduction de contraintes temporelles et le recours à la NMF semi-supervisée :

- **Contraintes temporelles :** La mise à jour des activateurs temporels dans la NMF se fait par défaut trame par trame, sans prendre en compte la corrélation qu'il peut y avoir entre la trame n et les trames précédentes (voir relation (2)). Néanmoins, la plupart des sons réels ont une évolution temporelle lente, notamment au sein d'environnements sonores urbains où la durée des sons (notamment le trafic routier) est de plusieurs secondes. Prendre en compte les trames temporelles précédentes permettrait donc que les formes des activateurs soient plus réalistes. Ceci est assuré en pratique en appliquant une pénalité $C_t(H)$ au calcul de la distance à minimiser dans la relation (3) [21].
- **NMF semi-supervisée :** la version de la NMF appliquée ici est celle de la NMF supervisée, pour laquelle le dictionnaire **W** est entièrement connu. Mais, elle peut être mal adaptée aux multiples environnements sonores. Dans la NMF semi-supervisée (abrégé NMF-SS [22], on considère un dictionnaire **W**, de dimension F x (M + J) dont une partie du dictionnaire est connue, $\mathbf{W_s}$, de dimension F x M et une seconde partie inconnue, $\mathbf{W_r}$, de dimension, F x J avec J << M. Lors de l'exécution de la NMF-SS, la partie inconnue sera apprise en même temps que l'intégralité de la matrice **H** (de dimension (M+J) x N). En pratique, une telle modélisation permet d'adapter la méthode à des mixtures sonores présentant des sons inconnus aux spectres caractéristiques, tels que des bruits de klaxons, de tondeuses, de sirènes, de cloches d'église…

	Parc		Rue calme		Rue animée		Rue très animée	
	NMF	**Filtre**	**NMF**	**Filtre**	**NMF**	**Filtre**	**NMF**	**Filtre**
f_c	1000	1000	2000	1000	5000	1000	5000	1000
W	Trafic	/	Trafic	/	Trafic	/	Trafic	/
$RMSE_{min}$	**3,28**	5,85	**1,33**	1,90	0,51	**0,43**	0,64	**0,63**

Tabl. 3 : Erreur RMSE minimale et sa combinaison de paramètres pour chaque environnement sonore

Conclusion

Cette étude propose l'utilisation de la méthode de la factorisation en matrice non-négative afin de de déterminer les niveaux sonores du trafic routier à partir d'enregistrements sonores réalisés en ville. L'application de cette méthode sur des scènes sonores simulées, dont le réalisme a été validé par un test perceptif, révèle que la NMF propose de meilleures estimations du niveau sonore que de simples filtres fréquentiels, notamment dans les environnements sonores peu ou faiblement impactés par les bruits du trafic. Des recherches sont en cours qui amélioreront vraisemblablement encore les estimations dans un avenir proche. A terme, la méthode pourra être étendue à d'autres sources de bruit, permettant non seulement une meilleure comparaison entre les cartes de bruit de trafic produites et les mesures, mais ouvrant également la porte à la production de cartes de bruit déclinées par types de sources. L'outil de reconnaissance des sources proposé offrira alors une certaine plus-value aux réseaux de capteurs développés actuellement en permettant une meilleure caractérisation des environnements sonores urbains.

Remerciements

Jean-Rémy Gloaguen est un doctorant co-financé par l'Ifsttar et la région Pays-de-la-Loire. Les auteurs tiennent à remercier Catherine Lavandier, pour le partage des données Grafic.

Références bibliographiques

[1] Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002, relating to the assessment and management of environmental noise. Official Journal of the European Communities. 2002.

[2] CERTU, "Comment réaliser les cartes de bruit stratégiques en agglomération ? Mettre en œuvre la directive 2002/49/CE », 2007, RF05807, janvier 2007.

[3] SETRA, « Prévision du bruit routier. 1. Calcul des émissions sonores dues au trafic routier.

[4] SETRA, « Prévision du bruit routier. 2. Méthode de calcul de propagation du bruit incluant les effets météorologiques (NMPB 2008).

[5] D. Oldoni, B. De Coensel, M. Boes, M. Rademaker, B. De Baets, T. Van Renterghem, and D. Botteldooren, "A computational model of auditory attention for use in soundscape research," J. Acoust. Soc. Am. 134(1), 852–861 (2013).

[6] A. Can A, C. Lavandier, P. Delaitre, L. Brocolini. Vers une meilleure description des environnements sonores urbains grâce aux mesures mobiles. Acoustique et technique, n°77, p39-48.

[7] J.-J. Aucouturier, B. Defreville, F. Pachet. The bag-of-frames approach to audio pattern recognition : A sufficient model for urban soundscapes but not for polyphonic music. The Journal of the Acoustical Society of America, vol. 122, no.2, pp.881-891, 2007.

[8] P. Paatero, U. Tapper, Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, vol. 5 no. 2, pp. 111–126, 1994

[9] D.D Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, vol. 401, pp. 788-791, 1999

[10] D. Guillamet, J. Vitrià, B. Schiele, Introducing a weighted non-negative matrix factorization for image classification, Pattern Recognition Letters, vol. 24, no. 14, pp. 2447-2454, 2003

[11] W. Xu, X. Liu, Y. Gong, Document Clustering Based on Non-negative Matrix Factorization, Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, pp. 267-273, 2003

[12] P. Smaragdis, J.C. Brown, Non-negative matrix factorization for polyphonic music transcription, Applications of Signal Processing to Audio and Acoustics workshop, IEEE Workshop on, New Paltz, NY, USA, 19-22 October 2003

[13] N. Bertin, Les factorisations en matrices non-négatives : approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique, Paris, Télécom ParisTech, 2009

[14] J.-R. Gloaguen, A. Can, M. Lagrange, J.-F. Petiot. Estimating traffic noise levels using acoustic monitoring : a preliminary study. DCASE2016, Detection and Classification of Acoustic Scenes and Events workshop, IEEE AASP Challenge, Budapest, Hungary, 03 September 2016

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

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

[17] J. Salamon, C. Jacoby, and J. Bello. A Dataset and Taxonomy for Urban Sound Research. Proceedings of the ACM International Conference on Multimedia

[18] P. Aumond, A. Can, B. De Coensel, D. Botteldooren, C. Ribeiro, & C. Lavandier. Modeling Soundscape Pleasantness Using perceptual Assessments and Acoustic Measurements Along Paths in Urban Context. Acta Acustica United with Acustica, 103(3), 430–443, 2017

[19] J.-R. Gloaguen, A. Can, M. Lagrange, and J.-F. Petiot, "Creation of a corpus of realistic urban sound scenes with controlled acoustic properties," The Journal of the Acoustical Society of America, vol. 141, no. 5, pp. 4044–4044, May 2017

[20] J. Pagès, E. Périnel. Blocs incomplets équilibrés versus plans optimaux. Journal de la Société Française de Statistique 148(2) :99–112, 2007

[21] S. Essid and C. Févotte. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. IEEE Transactions on Multimedia, 15(2) :415–425, Feb. 2013

[22] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "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, vol. E97.A, no. 5, pp. 1113–1118, 2014