Utilisation d'antennes à grand nombre de microphones pour la reconnaissance automatique de sources sonores en environnement urbain

Raphaël Leiba, François Ollivier, Jacques Marchal, Régis Marchiano
Sorbonne Universités
UPMC Université Paris 6
Institut Jean Le Rond d'Alembert
CNRS UMR 7190
4, place Jussieu
75005 Paris

Raphaël Leiba, Nicolas Misdariis
IRCAM-STMS-UPMC
CNRS UMR 9912
1, place Igor Stravinsky
75004 Paris
E-mail : raphael.leiba@upmc.fr

Résumé
La caractérisation de l'environnement sonore urbain est réalisée classiquement par la mesure d'indices énergétiques comme le niveau acoustique pondéré A (noté L_{Aeq}) ou le L_{den}. Cela a été réglementé par la Directive européenne 2002/49/EC qui impose aux grandes agglomérations la confection de carte de bruit en L_{den} pour l'information du public. Alors que ces mesures et simulations apportent une quantification moyenne des sources présentes dans l'espace sonore, notre approche s'inscrit dans l'estimation du paysage sonore, en apportant une quantification fine des sources sonores présentes.

Nous présentons ici une méthode de classification du trafic routier sur la base de catégories perceptives en utilisant un algorithme de machines à vecteurs de support (SVM). Une base de données d'apprentissage pour ce classifieur a été réalisée sur une piste d'essais avec différents types de véhicules à différentes allures. Une méthode d'antennerie acoustique est ensuite utilisée en milieu urbain pour extraire le signal des véhicules au passage. Ils sont isolés du trafic grâce au traitement des signaux d'antenne et vidéo. Les caractéristiques acoustiques sont alors extraites du signal et utilisées pour la classification. La robustesse de la classification est améliorée en ajoutant à la base d'apprentissage des signaux extraits des essais en ville. Finalement la répartition du trafic routier est présentée au long de la journée suivant les catégories perceptives établies par la littérature.

La directive européenne n°2002-49 du 25 juin 2002 relative à l'évaluation et à la gestion du bruit dans l'environnement a permis d'établir une réglementation normalisée. Cette directive impose aux agglomérations de plus de 100 000 habitants de mettre à disposition des citoyens des cartes du bruit issu des trafics routiers, ferroviaires ou aériens. L'indicateur utilisé pour constituer ces cartes est le niveau pondéré jour-soirée-nuit (L_{den}). Cet indicateur applique une majoration des mesures de 5dB en soirée et 10dB la nuit. Grâce à ces cartes, des actions de réduction des nuisances sonores ont d'ores et déjà été mises en œuvre. Cependant, Raymond Murray Schafer [1] estime qu'une bonne réglementation doit permettre d'agir sur les sources engendrant le plus de nuisances et pour cela se fonder sur une classification adaptée. La directive actuelle n'est pas orientée vers ce type d'information. En effet, elle ne considère pas la dimension perceptive et ne demande pas de distinguer, par exemple, les nuisances dues à chaque véhicule pour différents régimes de conduite.
Récemment en France, plusieurs consultations ont permis d'identifier le trafic routier comme l'une des nuisances sonores majeures. Des études ont tenté de mieux comprendre la gêne induite par le trafic routier à partir du signal audio de chaque source [2], [3].

S'inscrivant dans le contexte des travaux de R. M. Schafer, Morel *et al.* [4] mettent en avant la différence de perception pour différents types de véhicules routiers à différents régimes de passage. Il propose alors une typologie perceptive issue d'une tâche de catégorisation libre de signaux de véhicules au passage enregistrés dans la ville.

Cette typologie est déclinée en sept catégories qui mélangent parfois des véhicules de types différents dans les mêmes conditions de conduite, ou qui contiennent un seul type de véhicules dans une seule situation de conduite. Ces catégories sont recensées dans le tableau suivant:

#	Description
1	Deux-roues à vitesse constante
2	Deux-roues en accélération
3	Bus, véhicules lourds ou légers à vitesse constante
4	Deux-roues en décélération
5	Bus, véhicules lourds ou légers en accélération
6	Véhicules légers en accélération
7	Bus et poids lourds en accélération.

Tab. 1 : Catégories perceptives de véhicules de Morel et *al.* [4]

Caractériser le trafic routier à l'aide de cette typologie, faciliterait la construction de modèles de gêne et permettrait d'enrichir considérablement les cartes de bruit. C'est pourquoi notre travail propose un outil de classification automatique couplée à cette taxonomie perceptive du bruit de trafic routier.

Nous avons choisi de réaliser la classification sur des signaux audio spécifiques des sources prises individuellement, ce qui nécessite, d'une part, d'identifier les sources potentielles au sein du flux de véhicules et, d'autre part, d'extraire de la scène acoustique globale leur signal propre. Diverses méthodes d'extraction audio ont été proposées au cours des dernières années, principalement dans le domaine du traitement de la parole. Elles réalisent pour la plupart un débruitage de signaux monocanal, parfois en utilisant des algorithmes de NMF (*Non-Negative Matrix Factorization*) [5] ou des réseaux de neurones profonds [6], [7]. Nous avons privilégié une approche basée sur l'utilisation de grands réseaux de microphones et d'une méthode inverse standard adaptée aux sources en mouvement : la formation de voies conventionnelle [8], [9].

Cet article présente une méthode de classification des sources de bruits routiers qui combine une procédure de poursuite vidéo, l'utilisation d'un grand réseau de microphone pour réaliser le filtrage spatial des signaux des sources acoustiques et des algorithmes d'apprentissage machine et de classification des sources. Finalement la méthode décrite est appliquée en environnement urbain réel et une analyse quantifiée du trafic urbain sur une journée est présentée en fonction des catégories perceptives détaillées plus haut.

Méthodologie

Description générale

Ce paragraphe décrit la méthode de classification. Notre objectif est d'identifier le bruit produit par les véhicules pris isolément et en mouvement. La méthode procède en trois étapes. Tout d'abord la trajectoire des sources mobiles est déterminée par traitement d'images vidéo. Ensuite, en suivant sa trajectoire, on extrait le signal spécifique de chaque source à l'aide d'un algorithme de formation de voies standard adapté aux sources mobiles. Enfin, des descripteurs physiques ou psychoacoustiques extraits de ces signaux sources alimentent un étage de classification basé sur l'apprentissage machine pour affecter chaque source à sa catégorie perceptive.

Détermination de la trajectoire par *tracking* vidéo

Une scène de trafic routier typique implique divers véhicules qui suivent chacun une trajectoire différente. Cette première phase cherche à isoler chaque véhicule automatiquement par traitement du signal vidéo issu d'une webcam disposée au centre de l'antenne acoustique (voir paragraphe suivant). Le traitement vidéo procède image par image. Il est basé sur la soustraction d'arrière-plan et s'appuie sur *OpenCV*[1] suivant 4 étapes :

- L'arrière-plan est identifié lorsqu'aucun véhicule ne figure dans la scène filmée.
- Cette image constante et celles à traiter sont converties en niveaux de gris et floutées.
- La différence entre les 2 images produit une image «delta» floutée à laquelle est appliqué un seuil de niveau.
- Une détection de contour permet d'isoler les véhicules mobiles à l'intérieur de rectangles dont le centre décrit la trajectoire de la source considérée

La trajectoire de la cible est déduite des positions successives du centroïde pour aboutir à la position $\mathbf{x_i}$ de la cible mobile à chaque instant.

Une méthode d'imagerie pour extraire le signal source

Une fois la trajectoire connue, le signal de la source peut être extrait. Ceci est réalisé à l'aide d'une méthode inverse largement répandue : la formation de voies (« *beamforming* »). Avec cette méthode, le signal audio $s_i(t)$ associé à un véhicule en mouvement est estimé en appliquant aux signaux de l'antenne acoustique des décalages de phase de façon à pointer le faisceau selon la trajectoire du centroïde.

L'opération de formation de voies à une fréquence donnée est définie par l'expression suivante :

$$\hat{s}_i(\omega) = \frac{1}{N_m} \sum_m r_{mi} e^{\frac{j\omega r_{mi}}{c_0}} \cdot \hat{p}_m(\omega),$$

où $\hat{s}_i(\omega)$ est le signal du véhicule en mouvement à la pulsation ω, $r_{mi} = |\mathbf{x_m} - \mathbf{x_i}|$, la distance variable du microphone à la source poursuivie, N_m le nombre de microphones et $\hat{p}_m$ la pression enregistrée par le microphone m à la pulsation ω. Précisons que l'on peut extraire d'une même scène plusieurs sources en mouvement. La pression $p_m(\omega)$ à un instant donné est obtenue par transformée de Fourier à court terme (TFCT) d'une trame du signal temporel $p_m(t)$. Finalement ces opérations peuvent s'écrire de façon synthétique sous forme matricielle:

$$\hat{\mathbf{s}}_i(\omega) = \frac{1}{N_m} \mathbf{A}^H \hat{\mathbf{p}}_m,$$

où ^{H} marque la transposition hermitienne, $\hat{\mathbf{p}}_m$ le vecteur pression acoustique, et $\mathbf{A}$ une matrice $N_m \times N_i$ de décalage des phases avec, N_i le nombre de sources de la scène sonore.

1- *http://opencv.org*

Les opérations du domaine spectral sont parallélisées sur un processeur GPU [10] pour être exécutées en temps réel. Les signaux des microphones sont traités par trames de 50ms avec un recouvrement de 75%. Préalablement à la TFCT, les trames sont pondérées par une fenêtre qui assure la conservation de l'énergie dans les parties recouvertes. A la fin du traitement, le signal temporel associé à la source mobile isolée est obtenu par transformée de Fourier inverse.

L'algorithme de classification des sources

Dans cette partie, nous présentons la méthode de classification automatique utilisée pour catégoriser les véhicules à partir du signal spécifique obtenu à l'étape précédente. L'algorithme employé, le *Support Vector Machine* (SVM) permet une classification grâce à un apprentissage supervisé. Il cherche à affecter les caractéristiques extraites des signaux à leurs catégories en identifiant les limites séparant les catégories. Il est couramment utilisée pour la classification des signaux audio [11].

La Figure 1 présente un exemple de catégorisation de données. Ces données sont caractérisées par les variables x_1 et x_2 qui décrivent un espace plan. On observe deux catégories bien distinctes et séparées par une frontière clairement établie. Cette limite appelée « hyperplan » est orientée par le vecteur normal $\boldsymbol{w}$. L'algorithme SVM procède à la minimisation de la norme $||\boldsymbol{w}||$ pour rendre maximum la distance entre les deux catégories bornées chacune par les marges positionnées en $\boldsymbol{w} \cdot \boldsymbol{x} - b = +1$ et $\boldsymbol{w} \cdot \boldsymbol{x} - b = -1$. Les trois vecteurs $\boldsymbol{x}_i = (x_{1i}, x_{2i})$ délimitant ces deux plans sont appelés vecteurs de support. Le terme de vecteur est d'autant plus pertinent que les dimensions des vecteurs $\boldsymbol{w}$ et $\boldsymbol{x}$ sont grandes, c'est-à-dire quand le nombre de caractéristiques considéré est grand.

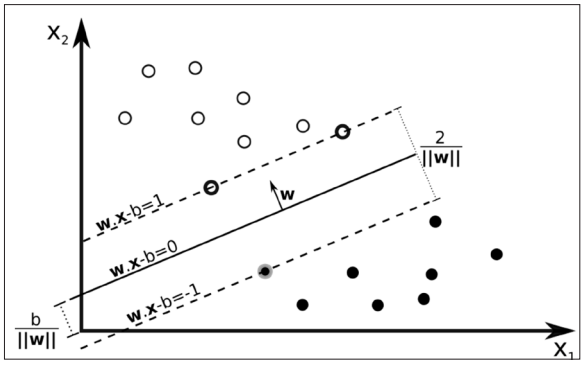

Fig. 1 : Exemple de catégorisation par SVM suivant 2 caractéristiques. Hyperplan de marge maximum (trait plein) et marges (droites en pointillés) dans le cas d'échantillons issus de 2 classes. Les trois échantillons situés sur la limite sont appelés vecteurs support

Valero *et al.*[12] ont comparé diverses caractéristiques extraites de signaux et ont mis en évidence la pertinence des coefficients cepstraux sur l'échelle de Mel[2] (*MFCC : Mel-Frequency Cepstral Coefficients*). Ces coefficients (décrits plus loin) sont largement utilisés dans le domaine de la reconnaissance vocale où, lorsqu'ils sont associés aux SVM, ils permettent d'obtenir des classifications avec de bons taux de réussite, [11], [13].

D'autres indicateurs courant sont été testés, physiques (comme le L_{eq} ou du centre de gravité spectral) et psychoacoustiques (comme la rugosité ou la sonie), mais ils ont fourni des résultats de moindre qualité.

Les MFCCs sont obtenus par filtrage de la densité spectrale de puissance (DSP) suivant un banc de filtres qui reproduisent, de façon simplifiée, la physiologie de l'oreille au niveau de la cochlée [14]. Les spectres issus des filtres sont combinés avant de subir une transformation pour produire les MFCCs. Dans notre étude cette suite d'opérations a été réalisée avec la librairie *python_speech_features*[3] pour produire finalement 13 MFCCs. Après normalisation ces coefficients servent de base à l'étape de classification. Dans cette dernière étape, l'algorithme SVM opère la classification suivant les 7 catégories listées plus haut. Le processus fait appel à la classe «C-Support Vector Classification» du module SVM de la librairie *python Scikit-learn*[4].

Megamicros : système de captation du champ acoustique

Le système *Megamicros* permet l'enregistrement simultané des signaux issus d'un grand nombre de microphones. Il est composé d'une antenne de microphones, et d'un module de pilotage et de multiplexage des signaux. L'ensemble est contrôlé par un ordinateur via une liaison série USB. L'élément de base de *Megamicros* est le microphone MEMS (*Micro Electro-Mechanical System*) numérique intégré dans un circuit électronique de quelques millimètres (voir Figure 2). Le conditionnement et la numérisation sont intégrés dans le circuit portant le microphone. Ceci présente l'avantage de réduire fortement l'encombrement et la connectique et d'alléger considérablement le déploiement des micros en grand nombre. De plus les caractéristiques de ces microphones numériques, sont parfaitement adaptées aux applications d'imagerie acoustique [10]. La production industrielle de ces composants de la téléphonie mobile contribue aussi à réduire fortement le coût de leur utilisation massive.

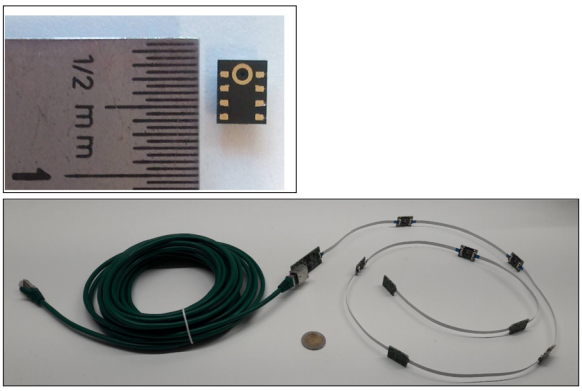

Fig. 2 : (en haut) microphone MEMS (ADMP441 de chez Analog Devices). (en bas) Faisceau composé de 8 microphones et de son câble.

2- L'échelle de Mel est une échelle psychoacoustique de hauteurs des sons, de grave et aigu, dont l'unité est le mel.
3- https://github.com/jameslyons/python_speech_features
4- http://scikit-learn.org

Par conception le système *Megamicros* permet de disposer de nombreux microphones sur de grandes étendues, de l'ordre de quelques dizaines de mètres suivant des géométries adaptées aux situations étudiées. Un faisceau unitaire du système regroupe 8 microphones numériques (Figure 2) qui peut avoir jusqu'à 3,5m de longueur. Le faisceau est relié à l'unité principale par un câble (type RJ45) dont la longueur peut atteindre 50 m sans perte de signal. L'unité principale du système permet de recueillir les signaux de 16 (128 microphones) ou 32 faisceaux (256 microphones). Elle est connectée à un ordinateur par liaison série USB3 ou éventuellement déportée à une très grande distance grâce à une fibre optique. Le système permet d'acquérir également quatre voies analogiques pour recueillir les signaux synchronisés d'autres capteurs pouvant servir de référence.
Enfin, des interfaces d'acquisition ont été développées sous Windows, macOs et Linux pour enregistrer et sauvegarder aisément l'ensemble des signaux. Les traitements de ces données brutes font appel à des algorithmes adaptés à chaque situation.
Megamicros est donc un système flexible, de faible coût, simple d'interfaçage. Il permet à ce jour l'acquisition synchrone de 256 microphones MEMS numériques avec une fréquence d'échantillonnage de 50 kHz pour couvrir la bande audible.

Véhicules isolés au passage : une base de donnée pour l'apprentissage

Pour valider la méthodologie proposée, une campagne de mesure a été menée afin de constituer une base de données de signaux issus de divers véhicules isolés passant devant une antenne suivant plusieurs allures.

Dispositif expérimental

Une antenne dédiée de 256 microphones a été déployée sur une piste d'essais. Cela permet de faire des mesures dans un environnement relativement contrôlé (bruit ambiant faible) et d'avoir une bonne maîtrise des véhicules lors des mesures (contrôle des trajectoires et des vitesses). L'antenne est rectangulaire, longue de 19,60 m et haute de 2,25 m. Elle est disposée suivant les normes de mesure de véhicules au passage : à 7,5 m parallèlement à la trajectoire des véhicules (voir Fig. 3).

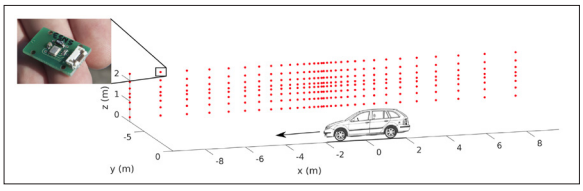

Fig. 3 : Schéma du dispositif expérimental utilisé pour les expériences de mesures de bruit au passage de véhicules. L'antenne contient 256 microphones (points rouges) MEMS (voir encadré) répartie. Adapté d'après [15].

Avec ce dispositif, 9 véhicules routiers (véhicule commercial lourd, véhicule léger et deux roues), dont le détail est donné en Tableau 2, ont été enregistrés à 3 allures différentes. Il faut noter que toutes les catégories définies par Morel et *al.* [4] figurent dans ce corpus.

Énergie	Cylindrée	Gamme
diesel	4 cyl.	berline
essence	3 cyl.	citadine
électrique		citadine
diesel	4 cyl.	monospace
diesel	4 cyl.	berline
diesel	4 cyl.	utilitaire
électrique		scooter
essence	50 cm^3, 2 temps	scooter
essence	400 cm^3, 4 temps	scooter

Tabl. 2 : Caractéristiques des véhicules utilisés lors des essais sur piste

La tâche d'extraction du signal d'un véhicule est d'abord testée sur le signal émis par un haut-parleur porté par une voiture au passage à vitesse constante devant l'antenne. Il s'agit d'un signal harmonique à 2 kHz. La Figure 4 présente le spectre d'un des micros de l'antenne (en bleu) et le spectre issu du *beamforming* (en orange). Les composantes tonales (fondamentale et harmoniques) sont parfaitement restituées et le bruit de fond global (bruit numérique des MEMS et bruit du véhicule) est réduit de 10 dB au-dessus de 1400 Hz, ce qui suggère que le processus de formation de voies est efficace dans le support spectral des sources.

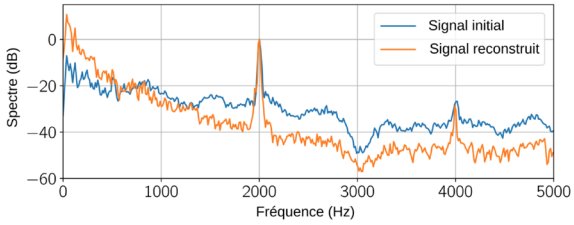

Fig. 4 : Comparaison des spectres d'un micro unique (en bleu) et de celui obtenu par formation de voies (en orange).

Résultats de classification

Des tests préliminaires ont montré que la pondération des MFCC améliore la classification. L'estimation de l'erreur est faite en entrainant le modèle sur 88% des données disponibles (68 signaux) et en le testant sur les données restantes. Les allures obtenues par le traitement vidéo sont prises en compte dans les données d'apprentissage. Trois informations sont retenues : véhicule en vitesse constante (=1 ou 0), en accélération (=1 ou 0) ou en décélération (=1 ou 0). Suite à nos tests préliminaires des vecteurs de pondération ont permis d'obtenir 100% de bonnes prédictions, l'un d'eux était :

$$Pondération = [\underbrace{2,1,0,4,2,1,1,0,1,1,1,1,}_{MFCC}\ \underbrace{1,1,1}_{Allures}].$$

On peut voir que certains MFCC ne participent pas au modèle (ici le 3 et le 8). Il nous est apparu que les MFCC n°3, 6, 8 et 11 peuvent même, dans certains cas, dégrader la classification.

Une étude paramétrique a permis de déterminer une pondération optimale des MFCC. La classification a été testée pour chaque combinaison de MFCC en affectant à chacun d'eux un coefficient entre 0 et 5 pour 8 MFCC (les douze premiers auxquels ont été retirés les MFCC n°3, 6, 8 et 11). Ainsi, 1 679 616 combinaisons sont possibles, mais seules 26301 permettent d'obtenir un score de 100% de reconnaissance.

Ces tests permettent de valider la procédure d'apprentissage et de reconnaissance. Le nombre élevé de combinaisons de MFCC permettant la reconnaissance est un élément positif qui apporte de la robustesse pour la reconnaissance dans un environnement urbain moins contrôlé.

Application en milieu urbain

La méthode de reconnaissance a été utilisée dans l'environnement urbain. Les résultats présentés ici ont été obtenus le long d'un axe comprenant 3x1 voies. L'enregistrement des signaux sonores utilisait une antenne spécialement conçue pour être déployée facilement en ville.

Dispositif expérimental

L'antenne utilisée est présentée Figure 5. Il s'agit d'un réseau linéaire de 128 microphones, long de 22m et placée sur un balcon surplombant la chaussée à 9m de hauteur. Une caméra est placée au centre de l'antenne. Le dispositif est séparé de la route par une rangée d'arbres (ces expériences ont été réalisées en hiver, les arbres sont donc dépourvus de feuilles). Le dispositif a été placé non loin d'une intersection avec des feux tricolores de signalisation (cf Figure 5-c) régulant le trafic. Un sonomètre et un anémomètre sonique ont également été incorporés au dispositif afin de s'assurer de la qualité des mesures.

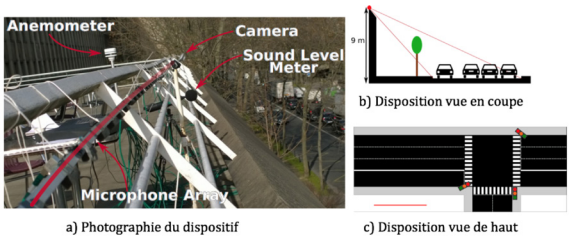

Fig. 5 : Visualisations du dispositif expérimental en environnement urbain réel

Une première étape de calibration a été réalisée à l'aide d'une source fixe contrôlée, disposée en différents points de la scène. La source émet un signal mono fréquentiel à 1kHz. La Figure 6 montre la densité spectrale de puissance mesurée par un microphone MEMS placé au centre de l'antenne (courbe bleue) et celle obtenue en appliquant l'algorithme d'extraction par formation de voies (courbe rouge) pour deux positions fixes du haut-parleur. Dans les deux cas, la fréquence d'émission est visible dans les spectres du microphone central même si l'émergence est faible par rapport au bruit de fond. Les spectres obtenus par la méthode d'extraction (courbes oranges) montrent une meilleure dynamique et une réduction du bruit à basse fréquence de l'ordre de 6dB.

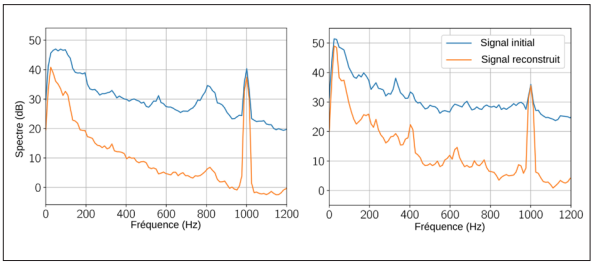

Fig. 6 : Densités spectrales de puissance mesurées par le microphone MEMS central et obtenues par la méthode d'extraction à deux positions du haut-parleur : (à gauche) 12,8m du centre de l'antenne, (à droite) 25,6 m de l'antenne.

Lorsque les sources sont mobiles, l'étape d'identification de la trajectoire est nécessaire. L'utilisation *in situ* a nécessité des adaptations mineures en raison du nombre important d'objets mobiles dans la scène. La Figure 7 montre le résultat du traitement vidéo à un instant donné. La partie (a) montre une image en noir et blanc où les parties blanches correspondent à des objets en mouvement (de bas en haut : une voiture, un scooter et un piéton). Sur cette image, on voit clairement l'effet de flou qui permet d'ignorer les taches diffuse s et fait correspondre aux objets réels (visibles dans (b)) une détection d'objet (cadres verts dans (b)). On reconstruit la trajectoire image par image (lignes rouges sur la Figure 7(b). On peut constater sur cet exemple que les trajectoires des véhicules sont bien reconstruites. Néanmoins, s'il y a trop de véhicules, il peut se produire un recouvrement des images floutées, ce qui dégrade les performances de la détection.

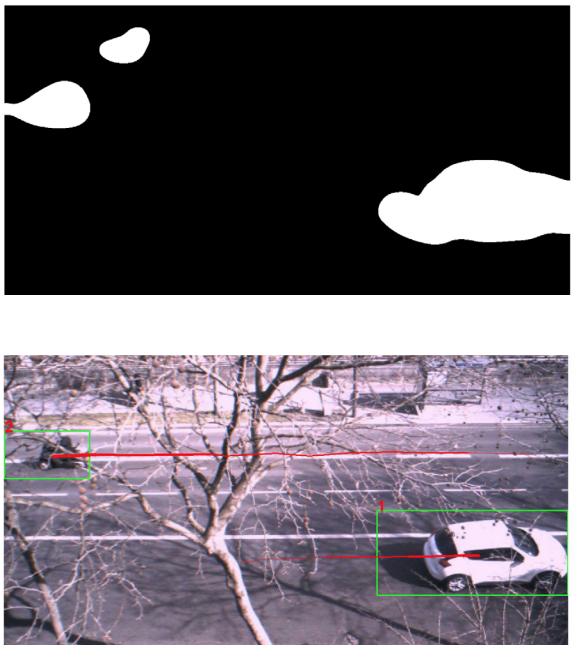

Fig. 7 : Illustrations de la procédure de traitement vidéo utilisée pour obtenir la trajectoire des véhicules : (a) image seuillée en noir et blanc des objets mobiles, (b) image originale à laquelle sont superposées les trajectoires (en rouge) et les objets mobiles détectés (rectangles verts).

La Figure 8 présente les spectrogrammes obtenus à partir du signal audio du microphone central (a) et à partir du signal extrait par formation de voies (b) d'un scooter passant devant l'antenne. Sur le spectrogramme (a), aucune fréquence particulière n'émerge du bruit large bande. En revanche, sur la partie (b) de la figure, les composantes tonales dues au moteur du deux-roues émergent clairement tout comme la partie large bande associé au contact pneu chaussée. Ce résultat illustre le gain obtenu par l'extraction du signal audio de chaque véhicule appartenant à la scène sonore. Ce gain doit permettre l'étape suivante de classification.

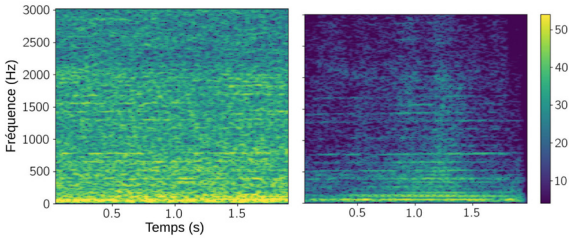

Fig. 8 : Spectrogrammes au passage d'un scooter calculés : (a) à partir du signal audio du microphone central, (b) à partir de la méthode d'extraction du signal par formation de voies.

Résultats de classification en milieu réel urbain

Dans cette partie, on applique la méthode complète jusqu'à la reconnaissance automatique des catégories définies par Morel et *al* [4] . Pour cela, on utilise les données enregistrées par le dispositif précédent afin d'alimenter l'algorithme d'apprentissage automatique. Les données ont été enregistrées au cours d'une même journée.

On teste l'algorithme de classification sur un segment de 10 minutes enregistré au milieu de la journée, traité pour affecter à chaque véhicule identifié une catégorie de façon supervisée. L'erreur d'identification atteint 68% avec le vecteur de pondération présenté précédemment. Il s'avère donc nécessaire d'enrichir la base de donnée d'apprentissage de données mesurées en situation réelle. Les dix minutes traitées permettent de constituer 186 échantillons d'apprentissage supplémentaires.

On évalue les performances du processus en calculant une matrice de confusion qui compare les catégories attendues (obtenues de façon supervisée) aux catégories estimées. Un cas de reconnaissance parfaite donnerait donc une matrice diagonale.

On calcule cette matrice sur la base de données étendue aux 186 échantillons, elle est présentée dans le Tableau 3. L'erreur globale est fortement réduite pour atteindre environ 14%, ce qui comparable aux résultats rapportés dans la littérature pour cette méthode [12]. Les erreurs sont commises, pour la plupart, sur la catégorie 1 (véhicule deux-roues à vitesse constante) et la catégorie 7 (véhicules lourds en accélération). Les nombres de véhicules classés dans les catégories 3 et 6 sont surestimés contrairement à toutes les autres catégories qui sont sous-estimées. À noter qu'aucun véhicule de la catégorie 4 n'était présent sur cette séquence.

		Estimée						
		Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Cat 6	Cat 7
Attendue	Cat 1	69,2	7,7	15,4	0	0	0	7,7
	Cat 2	0	75	16,7	0	0	8,3	0
	Cat 3	0,7	0	90,4	0	0	8,9	0
	Cat 4	0	0	0	100	0	0	0
	Cat 5	3,1	0	9,4	0	81,3	6,2	0
	Cat 6	0	2,9	5,7	0	5,7	85,7	0
	Cat 7	0	0	0	0	0	37,5	62,5
Somme		73,0	85,6	137,6	100	87,0	146,7	70,2

Tabl. 3 : Matrice de confusion pour les catégories perceptives (en pourcentage). La catégorie 4 n'est présente que pour les mesures sur poste d'essai.

On peut noter qu'avec ce dictionnaire d'apprentissage qui associe données issus de véhicules isolés et données extrait du trafic routier, l'erreur de classification, pour un test sur les passages isolés, remonte à 1,47%.

La même méthode est appliquée aux autres segments enregistrés. Les résultats sont synthétisés sur la Figure 9 qui présente l'évolution de la composition du trafic routier exprimée en catégories perceptives. On suppose, d'après les résultats précédents, que l'erreur est de l'ordre de 15%. La composition du trafic est assez stable au long de la période étudiée. On note vers 13h25, l'augmentation de la part des deux-roues en vitesse constante (catégorie 1) et la diminution des véhicules lourds et légers passant à vitesse constante (catégorie 3). La catégorie 4 est sous-estimée car la mesure de la trajectoire des véhicules deux roues est difficile quand ils sont masqués par d'autres véhicules.

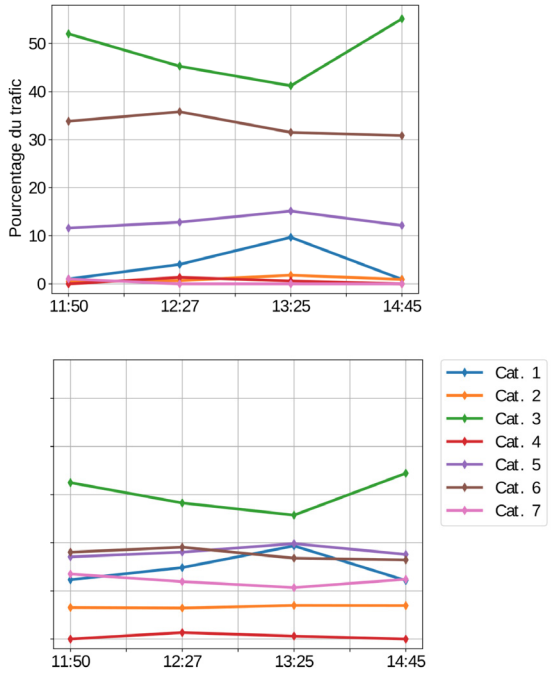

Fig. 9 : Evolution de la composition du trafic routier en pourcentage de chaque catégories perceptives présentes à différentes heures.

Ces résultats dépendent de la composition du trafic, donc des spécificités de l'emplacement choisi et ne sont pas généralisables. Comme illustré à la Figure 5-c, le dispositif est placé non loin d'une intersection. Les flux de véhicules sur la partie à trois voies et celle à une voie sont donc très différents : essentiellement en accélération pour les véhicules provenant de l'intersection et en décélération pour les véhicules allant vers l'intersection.

Enfin la Figure 9-b présente une correction de la composition du trafic routier en catégories perceptives. Cette correction a été faite à partir de la matrice de confusion en supposant que les erreurs restent les mêmes. La première conséquence est que la catégorie 5 devient aussi importante que la 6. Les catégories 1 et 7 sont aussi rehaussées significativement.

La stabilité de la composition du trafic est raisonnable compte tenu de la plage horaire (11h50 à 14h50) pendant laquelle les mesures ont été faites. L'erreur globale de 15% nous semble acceptable dans cette configuration exigeante (plusieurs voies dans les 2 sens et proches d'une intersection).

Conclusions

Cette étude a permis de mettre au point une méthodologie de reconnaissance automatique du trafic routier classé suivant des catégories perceptives préalablement identifiées par des études psychoacoustiques. Pour cela, un grand réseau de microphones est associé à du traitement vidéo et à un algorithme de classification supervisé.

La technique d'imagerie par formation de voies adaptée aux sources mobiles est appliquée pour extraire le signal spécifique des sources isolées du trafic routier. Des caractéristiques audio (MFCCs) extraites des signaux permettent finalement la catégorisation automatique des véhicules isolés du trafic dans des conditions réelles.

La méthode de catégorisation est mise en place grâce à une méthode de machine learning supervisé (SVM). Les éléments du dictionnaire d'apprentissage sont les MFCCs et les informations d'allure des véhicules. Cette méthode de catégorisation a été validée par des essais en environnement contrôlé et adaptée pour mesurer l'évolution du trafic sur une grande artère parisienne.

On parvient à une description fine des scènes sonores urbaines qui peut constituer un outil d'analyse du trafic urbain à destination des pouvoirs publics. Ce travail constitue une première étape dans la perspective d'estimer la gêne sonore ressentie par les riverains. Dans la suite on fera appel à des modèles psychoacoustiques exploitant les catégories et les signaux obtenus par la méthode proposée.

Notons enfin que comme la méthode repose sur l'utilisation de vidéo et de microphones, l'équipement mis en oeuvre peut aussi être utilisé à d'autres fins comme la surveillance (identification et localisation des sources) ou la mesure du niveau sonore pour la cartographie.

Remerciements

Les auteurs remercient chaleureusement : Dominique Busquet (UPMC), Pascal Challande (UPMC), Jean-Christophe Chamard (PSA), Hélène Moingeon (UPMC), Christian Ollivon (UPMC) et Vincent Roussarie (PSA).

Cette recherche bénéficie du support de la Chaire «Mobilité et qualité de vie en milieux urbains», portée par la Fondation UPMC et soutenue par les Mécènes (PSA Peugeot-Citroën et RENAULT)

Références bibliographiques

[1] R. M. Schafer, Le Paysage Sonore, 4ème. Wildproject, 2010.

[2] A. Klein, C. Marquis-Favre, R. Weber, and A. Trollé, "Spectral and modulation indices for annoyance-relevant features of urban road single-vehicle pass-by noises," J. Acoust. Soc. Am., vol. 137, no. 3, pp. 1238–1250, Mar. 2015.

[3] J. Morel, C. Marquis-Favre, and L.-A. Gille, "Noise annoyance assessment of various urban road vehicle pass-by noises in isolation and combined with industrial noise: A laboratory study," Appl. Acoust., vol. 101, pp. 47–57, Jan. 2016.

[4] J. Morel, C. Marquis-Favre, D. Dubois, and M. Pierrette, "Road Traffic in Urban Areas: A Perceptual and Cognitive Typology of Pass-By Noises," Acta Acust. united with Acust., vol. 98, no. 1, pp. 166–178, Jan. 2012.

[5] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," {IEEE} Trans. Audio, Speech, Lang. Process., vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[6] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," Comput. Speech {&} Lang., 2016.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," {IEEE}/{ACM} Trans. Audio, Speech, Lang. Process., vol. 23, no. 1, pp. 7–19, Jan. 2015.

[8] J. E. Adcock, "Optimal filtering and speech recognition with microphone arrays," Brown University, 2001.

[9] I. Hafizovic, C.-I. C. Nilsen, M. Kjølerbakken, and V. Jahr, "Design and implementation of a {MEMS} microphone array system for real-time speech acquisition," Appl. Acoust., vol. 73, no. 2, pp. 132–143, 2012.

[10] C. Vanwynsberghe, R. Marchiano, F. Ollivier, P. Challande, H. Moingeon, and J. Marchal, "Design and implementation of a multi-octave-band audio camera for realtime diagnosis," Appl. Acoust., vol. 89, pp. 281–287, 2015.

[11] C.-C. Lin, S.-H. Chen, T.-K. Truong, and Y. Chang, "Audio classification and categorization based on wavelets and support vector Machine," {IEEE} Trans. Speech Audio Process., vol. 13, no. 5, pp. 644–651, Sep. 2005.

[12] X. Valero and F. Ali'as, "Automatic classification of road vehicles considering their pass-by acoustic signature," in ICA, 2013.

[13] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using One-Class {SVMs} and Wavelets for Audio Surveillance," {IEEE} Trans. Inf. Forensics Secur., vol. 3, no. 4, pp. 763–775, 2008.

[14] H. Fastl and E. Zwicker, Psychoacoustics - Facts and Models, 3rd ed. Berlin : Springer Verlag, 2007.

[15] R. Leiba, F. Ollivier, R. Marchiano, N. Misdariis, and J. Marchal, "Urban acoustic imaging : from measurement to the soundscape perception evaluation," in INTER-NOISE 2016, 2016.