# Acoustical Classification of the Urban Road Traffic with Large Arrays of Microphones

Raphaël Leiba, François Ollivier, Régis Marchiano, Nicolas Misdariis, Jacques Marchal, Pascal Challande

**Acta Acustica united with Acustica**
**Vol. 105 (2019) 1067 – 1077**

DOI 10.3813/AAA.919385

# Acoustical Classification of the Urban Road Traffic with Large Arrays of Microphones

Raphaël Leiba[1,2)], François Ollivier[1)], Régis Marchiano[1)], Nicolas Misdariis[2)], Jacques Marchal[1)], Pascal Challande[1)]

[1)] Sorbonne Université, CNRS, Institut Jean Le Rond ∂'Alembert, 75005 Paris, France.
raphael.leiba@sorbonne-universite.fr

[2)] STMS Ircam-CNRS-SU

*Colour Figures: Figures in colour are given in the online version*

**Summary**

This work is part of a study dealing with city-dwellers' quality of life. Noise is known to be an important factor influencing the quality of life. In order to diagnose it properly, we propose a noise monitoring system of urban areas. It is based on the use of large microphone arrays in order to extract the radiated sound field from each passing-by vehicle in typical urban scenes. A machine learning algorithm is trained so as to classify these extracted signals in clusters combining both the vehicle type and the driving conditions. This system makes it possible to monitor the evolution of the noise levels for each cluster. The proposed system was first tested on passing-by isolated vehicles measurements and then implemented in a real street in Paris (France).



**PACS no. 43.50. 43.50.Lj, 43.50.Rq, 43.60.Yw, 43.60.Bf, 43.60.Cg, 43.60.Fg, 43.60.Jn, 43.60.Np, 43.60.Vx**

## 1. Introduction

Quality of life is a major challenge in urban places because of the large number of environmental factors influencing it. In projects such as Mouvie [1] systemic approaches are needed to work on both air pollution and sound pollution and their impact on the comfort of city-dwellers.

Notably, noise has stress-related impacts on human health such as sleep disturbances or cardiovascular diseases [2, 3, 4]. The World Heath Organisation (WHO) indicates [5] that between 1 and 1.6 million years in good health are lost (disability-adjusted life-years, or DALYs) every year in occidental Europe because of transportation noise. Sleep disturbance is the major effect with 903 000 years lost every years and long-term noise annoyance (due to passive effects of noise) is the second effect (654 000 years lost).

In order to unify the national initiatives and provide an efficient tool to diagnose the urban sonic environment, European parliament voted the 2002/49/CE directive in 2002 [6]. It imposes for the large cities of the member countries of the European union to realise noise maps based on the $L_{den}$ index (stands for Day-Evening-Night level). It aims to take into account the influence of time-periods of the day in noise annoyance by increasing by 5 dB(A) the evening levels and by 10 dB(A) the night levels. After this diagnosis step the large cities (over 100,000 inhabitants) have to provide action plans showing their ambition to reduce the proportion of persons highly annoyed by noise with practical actions.

These regulations, with regard to WHO's threshold values [7, 8], provide an efficient way to inform the city dwellers on their average exposure to noise. But these maps have some limits. First, they are provided for each type of source separately (road, aircraft, railway and industrial) whereas in urban areas city dwellers are usually exposed to several sound sources at the same time – with possible interactions and not just a summation. So that the overall exposure is not estimated. Second, whereas the variation of noise exposure over the day may be of interest for city-dwellers and urban planners, it is averaged because of the use of the $L_{den}$. Third, the noise maps are obtained by simulations and rarely confronted to measurements and consequently validated experimentally. Therefore, it is interesting to see the emergence of monophonic acoustic sensor networks (such as CENSE or DYNAMAP [9]) to provide additional real-time information about the sonic environment.

In this work, we focus on the noise induced by the road traffic. In this specific context, noise maps have additional weaknesses. Thus, though they are often cited as the most annoying sources [10, 11], the powered two-wheelers where considered as light vehicles [11]. Indeed, in many cities the traffic flow estimation is based on counting the number of axles of the passing-by vehicle, therefore it cannot distinguish a powered two-wheeler from a car.

Received 17 June 2019,
accepted 25 October 2019.

Note that the European Union voted in 2015 a directive [12] redefining the categories to be used for noise mapping, including the powered two-wheelers. The application of this directive had to be applied before 31/12/2018.

In addition, Marquis-Favre *et al.* [13] reported that the perceived noise annoyance induced by road traffic (here the short-term annoyance, estimated with active listening) is related to the mode of transportation (road, rail or air traffic) but also to the vehicle type. The driving conditions are also pointed out as a decisive factor to this annoyance. Thus it appears that the road traffic estimation should be improved to detect all type of vehicles and their driving conditions as well.

In this aim, we propose to capture the audio signal of each road vehicle, extract from it the vehicle type and its driving condition in order to provide a more detailed description of the road traffic noise and pave the way of short-term noise annoyance estimation in urban areas.

If the audio signal is known, it can be used to identify the vehicle. Indeed, sound source classification has been investigated with machine learning based on monophonic signals in the past decade (see *e.g.* [14, 15, 16, 17]). But in major streets the spatial and temporal masking effects of the different sound sources prevent from classifying each vehicle properly.

Multiple studies have been conducted to separate sound source on monophonic signals (see for example Gloaguen *et al.*[18] for recent development in Non-Negative Matrix Factorisation applied to urban sound scenes) but we assume that a spatial filtering technique will provide better results and extract each vehicle audio signal with more accuracy.

Weinstein *et al.* [19] showed that microphone arrays can be used for sound sources separation using inverse techniques. The application to low-speed moving source signal extraction has been done by Hafizovic *et al.*[20] on a basketball court with a 300 microphones array.

In this article, we propose a road traffic monitoring system. It aims at detecting each vehicle type, identifying its driving conditions and extracting its specific sound signal. This permits to compute indices that could be used to better assess noise annoyance such as loudness. By identifying the vehicles and isolating their audio features, the proposed system provides more detailed information than those provided by standard urban noise observatories.

Part 2 presents the tools implemented in the study. First a video tracking method provides the trajectory of each vehicle (see sec. 2.1). From this trajectory, the system uses large microphone arrays (sec. 2.2) together with a dedicated beamforming technique to extract the signal of each vehicle embedded in the traffic (sec. 2.3). The last step of the process consists in classifying these signals into clusters mixing both vehicle type and driving condition (sec. 2.4). Note that this study only focuses on internal combustion vehicles.

Part 3 presents the applications of the method. In a first step, it is used in a controlled set up to characterise isolated vehicles on a test track (Section 3.1) and constitute a learning database. In a second step, the system is implemented in a real urban context to evaluate its performances in terms of classification according to objective features (Section 3.2). Finally, the system is modified to perform classification according to perceptual indices in Section 3.3.

Finally, Part 4 exposes the global outcome of this work by presenting the evolution of the sound level all over a day with respect to the estimated perceptual clusters.

## 2. Materials and Methods

The method for an acoustical classification of the urban road traffic starts from a tracking step of each vehicle in the traffic flow (Section 2.1). This provides the trajectory of the sound sources to be measured and classified. The individual signal extraction relies on the implementation of large microphone arrays presented in Section 2.2. The method uses beamforming technique dedicated to moving sources and presented in Section 2.3. The last step aims at classifying the extracted signals. The method based on a supervised machine learning process is presented in Section 2.4.

### 2.1. Moving vehicle tracking method

In order to obtain the trajectory of each vehicle embedded in the traffic, we developed an in-house tracking method.

We first perform a contour detection on the video file recorded by a camera located at the centre of the microphone array. It is based on background subtraction for each video frame using the `OpenCV` library[1]. The background is created by averaging the 500 preceding frames. Each moving object is reduced to a rectangle including it. The rectangles connection between two consecutive frames is simply computed by finding the minimum distance between two rectangle centroids. Finally, the vehicle trajectory is obtained by gathering the connected centroids.

Figure 1a shows an example of vehicle tracking for a pass-by measurement. But the trajectory detection also has to be robust to the presence of obstacles (such as trees) between the camera and the vehicle, like in the configuration presented in Figure 1b: camera over a multi-lane street. To do so, the current frame – to whom the background is subtracted – is blurred in the direction of the vehicles.

### 2.2. Microphone arrays

#### 2.2.1. The *Megamicros* acquisition system

The *Megamicros* project, introduced by Vanwynsberghe *et al.*[21] with a 128 microphones array, aims at providing digital acquisition systems able to capture up to 1024 synchronised acoustic signals. These systems are dedicated to applications such as acoustic imaging, room acoustics or source directivity measurements. Based on digital MEMS microphones these systems are very versatile and easy

[1] Available at opencv.org

Table I. Characteristics of the vehicles for the track tests. `hv` stands for heavy vehicle, `lv` for light vehicle and `twv` for twho-wheeler vehicle.

| Label | Energy | Engine | Range |
|---|---|---|---|
| `lv1` | diesel | 4 cyl. | sedan |
| `lv2` | gasoline | 3 cyl. | urban |
| `lv3` | diesel | 4 cyl. | minivan |
| `lv4` | diesel | 4 cyl. | sedan |
| `hv` | diesel | 4 cyl. | 25 $m^3$ utility |
| `twv1` | gasoline | 50 $cm^3$, 2 stroke | 2-wheeler |
| `twv2` | gasoline | 400 $cm^3$, 4 stroke | 2-wheeler |

to set-up. The MEMS microphones (ADMP441 - Analog Device) are omnidirectional and have a rather flat frequency response between 60 and 15,000 Hz. These systems allow to build arrays of arbitrary geometries, possibly with extensions of a few tens of meters. Two of such arrays were implemented in this study. They are presented in Sections 2.2.2 and 2.2.3. In two following experiments, the signals are sampled at 50 kHz.

### 2.2.2. Isolated vehicle pass-by measurements

The microphone array used for the test-track experiment has been designed to provide the best possible acoustic image of the noise sources of passing-by vehicles [22]. Therefore, the microphone array is large enough to offer a sufficient resolution at low frequencies and dense enough to avoid grating lobes at high frequencies. The array was built according to the geometry presented in Figure 2. 256 microphones were distributed over a 20 m long and 2.25 m high area, thanks to 32 vertical uprights supporting 8 microphones each. The microphone array was located 7.5 m away from the vehicle path, following the ISO 362 recommendations for pass-by noise measurements. In detail, horizontally the inter-microphone distances are from 10 cm to 1.53 m and from 17 cm to 39.9 cm vertically.

We aimed at being representative of the urban road traffic in terms of vehicle types and driving conditions. Table I lists the characteristics of the vehicles involved in the track tests. Note that the vehicle named `hv` is the one considered as heavy vehicle despite it is a large utility vehicle, not a proper bus or truck. The study only focuses on internal combustion vehicles.

To simulate all possible urban driving configurations, each vehicle under test has passed the following scenarios in both back and forth ways:

- 25 km/h constant speed in second gear;
- 50 km/h constant speed in third gear;
- traffic light vicinity:
  - deceleration from 30 to 0 km/h;
  - 2 s stop, engine idling;
  - acceleration from 0 to 30 km/h with gear change;
- full throttle acceleration over 20 m.

### 2.2.3. Urban experiment

Our goal is to monitor the road traffic in urban environments. Therefore, *in situ* measurements have been carried

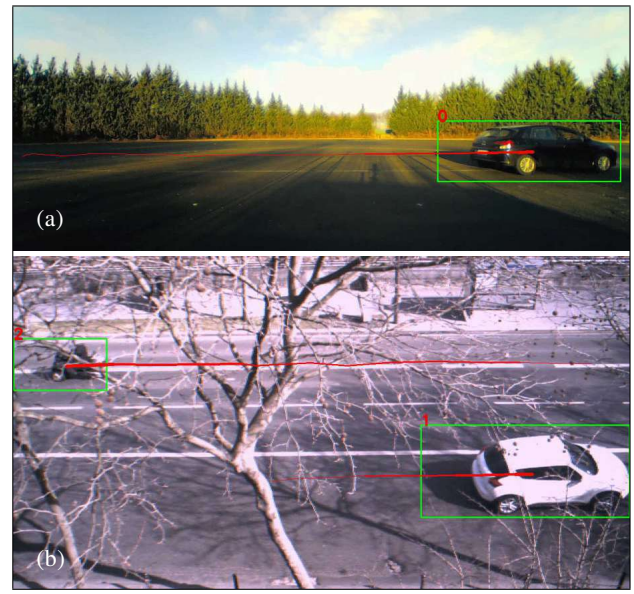


Figure 1. (Colour online) Detected moving vehicles (green rectangles) and trajectory extraction (red lines) for isolated vehicles or real urban situation. (a) Isolated vehicle on test track (test-track experiment), (b) Vehicles in a parisian street (*in situ* experiment).

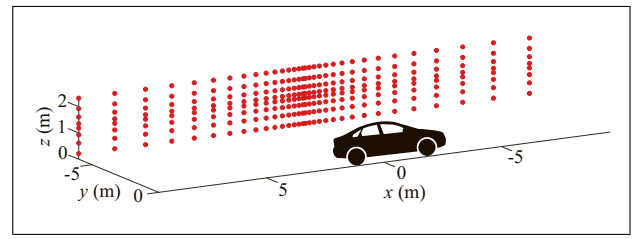


Figure 2. (Colour online) Microphone array positions (red dots).

out on a multi-lane urban street (St Bernard Quay in Paris - France) with a specific microphone array. It is 21.6 m long and is located 9.5 meters overhanging a 3×1 lane street at a 13.5 m distance from its centre (see Figure 3). It is composed of 128 MEMS microphones regularly spaced with a pitch of 17 cm. This configuration allows segregating the vehicles in the same lane in a wide frequency range, even at low frequencies thanks to the array length. The overhanging position of the array allows to have a phase difference between the lanes, which makes it possible to separate the sources located in different lanes. Due to its lightness and simplicity, the installation of the antenna only requires 30 minutes.

As illustrated in Figure 3, this experiment takes place in a street with an "L" shape, meaning that there is no building in the opposite side of the street. It is an important street and the microphone array is set-up close to a traffic light so that we can expect to have all the driving conditions described in the previous section.

This experiment took place during a day in winter which prevented the video tracking process to be disrupted by the leaves on the trees. Six sequences of ten minutes each have been recorded during day time.

### 2.3. Moving source signal extraction

Knowing the source position at all time, the microphone array recordings can undergo a beamforming (BF) process to extract audio signal of each passing-by vehicle from a multi-source sound scene.

In acoustics, BF is used as a reference method, since it is robust for source localization over a discretised plane or into a volume including static sound sources. The method can also be considered as a way of spatial filtering (see *e.g.* [19, 20]).

In this study, we propose to use this property on moving vehicles. Therefore, the standard free-field propagation model used in classical delay and sum (DAS) applications has been modified to take into account the cinematic of the vehicles. Figure 4 presents a classical scenario with a linear microphone array recording the sound field propagated by the $i^{th}$ monopolar sound source moving on a straight line.

Morse and Ingard [23], considering an homogeneous media and a free field, write the pressure at time $t_r$ at microphone $m$ emitted by source $i$ at time $t_e$ as

$$p_m(t_r) = \frac{s_i(t_e)}{r_{mi,e}\left(1 - M_a(t_e)\sin\theta_e\right)^2} \tag{1}$$

with $t_r = t_e + r_{mi,e}/c_0$, $r_{mi,e}$ the distance between the microphone $m$ and the source $i$ at emission time $t_e$, $s_i(t) = q\prime(t)/4\pi$, $q\prime(t)$ is the derivative of the source mass flow and $M_a(t_e)$ is the Mach number of the source at emission time $M_a(t_e) = V(t_e)/c_0$.

The reconstructed source signal $s_i(t_e)$ is estimated by Cousson *et al.* [24], for instance, with

$$s_i(t_e) = \frac{1}{N_m}\sum_{m=1}^{N_m} r_{mi,e}\left(1 - M_a(t_e)\sin\theta_e\right)^2 p_m(t_r) \tag{2}$$

with $N_m$ the number of microphones. The central term bears the only modification in the classical DAS expression.

However, in the rest of this paper the energy compensation is discarded (by removing the $r_{mi,e}$ factor) so that the output signal level is the one recorded by the microphones. Thus, the extracted signal writes:

$$\hat{s}_i(t_e) = \frac{1}{N_m}\sum_{m=1}^{N_m}\left(1 - M_a(t_e)\sin\theta_e\right)^2 p_m\left(t_r\right). \tag{3}$$

The signal $\hat{s}_i(t)$ is an estimator of what would have been recorded by one microphone if only the source $s_i$ where in the sound scene. By doing so, only a spatial filtering is operated and the results are more comparable to the noise maps (that provides $L_{den}$ in facade) and to the city dwellers feeling.

The simulations and the experimental tests presented in appendix A1 show that the BF method presented in equation (3) is accurate to extract each vehicle signal.

Note that in both experiments presented in this paper, the free-field model can be considered valid as there are no major reflectors except for the road, which is too close to the sources to have a real influence in these configurations.

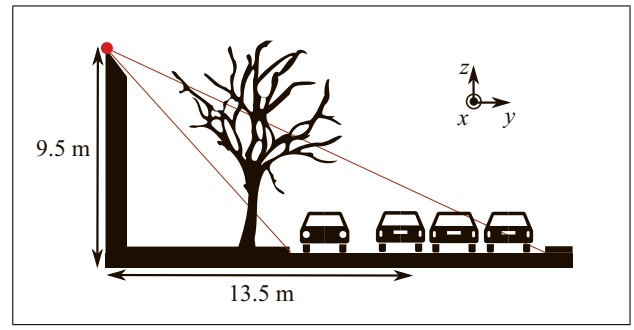


Figure 3. (Colour online) Scheme of the experiment configuration. The linear microphone array (red dot) is set-up over a 3×1 lane street in Paris.

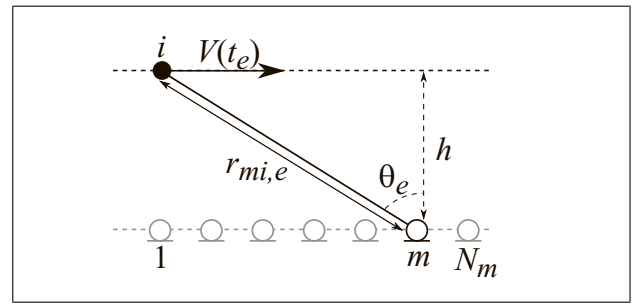


Figure 4. Configuration example with a $N_m$ linear microphone array and a $i^{th}$ sound source moving rectilinearly at a speed of $V$ at $t = t_e$.

### 2.4. Moving vehicle classification

In classification tasks, the challenge is not only in selecting the best algorithm but also in finding the best data to use: here the audio descriptors. Valero *et al.* [15] provides a comparison of classification accuracy of environmental noise obtained by 13 signal features and 4 different methods. They point out that the MFCCs (Mel Frequency Cepstral Coeficients) or the MPEG7 descriptors give good accuracy in noise source classification, especially for Gaussian Mixture Model (GMM), K-Nearest Neighbour (KNN) or Neural Network (NN).

MFCCs are widely used in noise source classification. As pointed out by Giannoulis *et al.* [25], a major part of the research teams uses MFCCs as audio descriptors for classifying different sound scenes signals. MFCCs, as presented by Davis and Mermelstein [26], result from different calculations over each signal frame, typically 25 ms long. First, the energy summation of the filtered spectrum over a triangle filter bank align along the mel scale (mimicking the cochlea) is taken. Then, the discrete cosine transform of the log of each energy is computed. For our purposes, MFCCs are calculated using the `python_speech_features` library[2].

During the DCASE 2013 challenge [25] the best results have been obtained by combining MFCCs with a Support Vector Machine (SVM). Introduced by Cortes and Vapnik [27], this method is widely used for supervised classification tasks. It aims at finding so-called hyper-plans that sep-

[2] Available at github.com/jameslyons/python_speech_features

Table II. Classification clusters.

| | 2-Wheeler | Light V. | Heavy V. |
|---|---|---|---|
| Const. Speed | 1 | 4 | 7 |
| Acceleration | 2 | 5 | 8 |
| Deceleration | 3 | 6 | 9 |

arate the samples of different classes with the maximum margin. The hyper-plans are defined by normal vector $\mathbf{w}$ and the margin width is equal to $2/\|\mathbf{w}\|$ so that minimising $\|\mathbf{w}\|$ maximises the margin. The $\mathbf{w}$ vector has the same number of components that the number of features used for classification (here the MFCCs). In order to allow errors and approximation in the case clusters could not be linearly separable, slack variables $\zeta_i$ – counting the number of errors – are added to relax the constraints on the learning vectors. Finding hyper-plans reduces at solving the equation

$$\hat{\mathbf{w}} = \underset{w,\zeta}{\operatorname{argmin}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathbf{C}\sum_{i=1}^{n}\zeta_i \,, \tag{4}$$

with $\mathbf{C}$ the parameter that determines the tradeoff between increasing the margin-size and ensuring that the samples lie on the correct side of the margin. $\mathbf{C}$ can be a vector or a scalar, providing respectively a value for each class or the same for all.

Note that convolutional neural network have also been used in environmental sound classification for the past years. This type of algorithm seems interesting, but so far, it gives the same type of performance as SVM but with an intensive computational cost (see *eg.* [28, 29]).

In our case, the task consists in classifying road vehicles in terms of type (2-wheeler, light vehicle and heavy vehicle) and driving conditions (constant speed, acceleration and deceleration). Nine clusters are used. They are detailed in Table II.

## 3. Implementation

The monitoring method proposed in the previous part is first tested and validated with isolated vehicles on a test track. Then, its application in a real urban scenario is presented.

### 3.1. Controlled set of vehicles

This extraction method has been applied to the pass-by measurements listed in Table I for the various driving conditions. Note that the "traffic light simulation" recordings are split into 3 driving conditions: deceleration, idle (not classified) and acceleration. Finally, 70 pass-by signals constitute the classification database (called test-track database in the rest of the article). They are distributed in the different clusters as presented in Figure 5.

Signal features selection is important for obtaining the best classification results. MFCCs are computed every 100 ms with 52 filters (ranging from 0 to 25 kHz, half the sample rate) and 26 MFCCs are obtained. A simplified

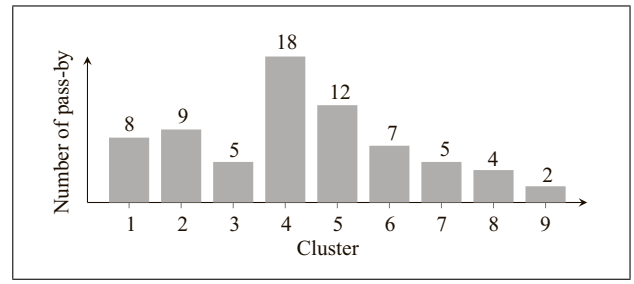


Figure 5. Number of pass-by measurements in each cluster.

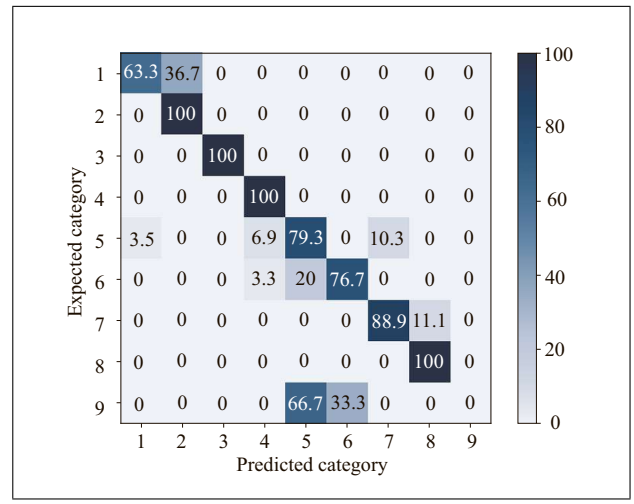


Figure 6. Confusion matrix in percentage of vehicle per category - Global classification accuracy: 88 %.

bag-of-frames approach [30] is used for representing the MFCCs evolution over the time: the time evolution is simplified to its mean and standard deviation. So that for each pass-by the extracted signal is represented by 52 audio features.

The classification is trained by SVM over 88 % of the signals and is tested on the remaining 12 % (8 signals). The results are averaged over 35 different learning and testing signal combinations.

Figure 6 presents the confusion matrix for this classifier. It shows some small confusion, mostly within the classes of a vehicle type. Some confusions outside a vehicle type are also present. For example, the light vehicle in acceleration (cat. 5) are classified 10 percent of the time as heavy vehicle in constant speed (cat.7) and 3.5 percent as a 2-wheeler in constant speed (cat.1)

Using the MFCCs as descriptors and a relaxation parameter $\mathbf{C}$ from 5 and above, the score is 88 % of correct classification which is good regarding the literature [15, 29, 28].

This result can be improved by adding more information to the dataset used by SVM: the driving conditions. Indeed, in this controled experiment they are well known. So that the learning and testing datasets are composed of 55 elements: 52 audio features and 3 binary values, one for each driving condition. Note that each element of the dataset is normalised by its maximum among the 70 pass-by data.

Table III. Number of pass-by measurements by category for training and testing datasets. The training dataset is based on test-track and *in situ* measurements.

| | Dataset | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Train | test-track | 8 | 9 | 5 | 18 | 12 | 7 | 5 | 4 | 2 | 494 |
| | *in situ* | 32 | 6 | 0 | 260 | 60 | 39 | 7 | 17 | 3 | |
| Test | *in situ* | 10 | 5 | 1 | 68 | 14 | 10 | 2 | 4 | 1 | 115 |

Figure 7 presents the confusion matrix for this classifier, the global classification accuracy rises to 99 %. It represents the percentage of predicted vehicle category with reference to the real (expected) one.

Only the heavy vehicle in deceleration (cluster 9) is misclassified, being considered as a light vehicle in acceleration 67 % of the time or in deceleration 33 % of the time. This could be explained by the nature of the vehicle: a large utility vehicle, powered by a car-like engine not a proper truck one. Note that tuning of $C$ parameter allows some samples to be misclassified. The counterpart of having a good fit of the SVM on the overall data is that the heavy vehicle in deceleration is 67 % of the time misclassified in acceleration.

### 3.2. Urban sound scene

From the *in situ* experiment presented in Section 2.2.3, six available recordings are distributed from 11:50 to 15:15. Three of them (11:50, 15:00 and 15:15) have been manually tagged in order to quantify the classification accuracy. It provides respectively 210, 246 and 83 vehicle trajectories (only the 200 first seconds have been tagged for the 15:15 recording) and forms what will be called the *in situ* database in the following.

The first classification tests have shown that we have to enhance the training dataset by adding data from the *in situ* database to the test-track database. The distribution between training and testing dataset is presented in Table III. Note that for some clusters the number of samples is very low. This is mainly due to the tracking method which is not robust enough but also to the low number of heavy vehicles passing during the measurements. For the third category, the two-wheelers trajectories were usually lost when they were overtaking another vehicle at idle. Note that the 52 audio features are normalised by the the same values (maximum of test-track database) and that the driving condition is deduced from the vehicle trajectory.

As discussed in Section 2.4, the $\mathbf{C}$ parameter can be a scalar (same value for all clusters) or a vector (one value per clusters). After a parametric study, the best classification results give 82 % of accurate classification. They are obtained for

$$\mathbf{C} = \left[6, 0.6, 0.6, 1.5, 1.5, 0.3, 6.6, 2.4, 3.3\right].$$

For more details, the confusion matrix is given in Figure 8. The good results are mainly due to the good classification of the light vehicles: from 70 to 93 % of accuracy. The

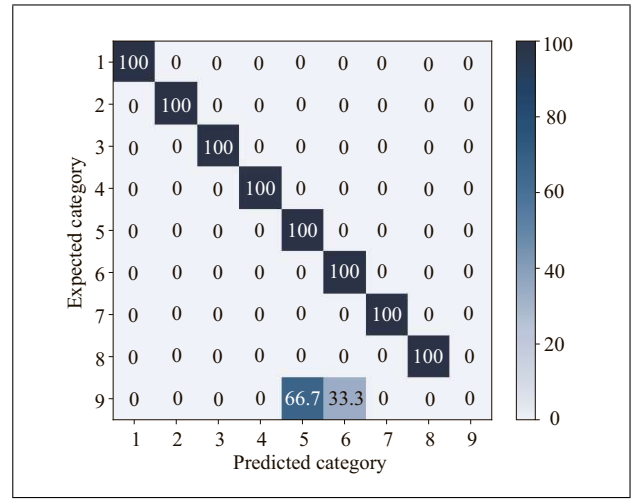


Figure 7. Confusion matrix in percentage of vehicle per category - Driving conditions added to the dataset - Global classification accuracy: 99 %.

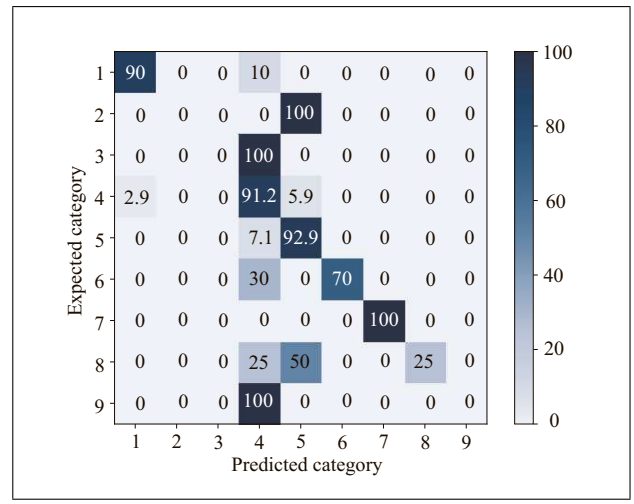


Figure 8. Confusion matrix in percentage of vehicle per cluster - Global classification accuracy: 82 %.

other clusters are often confused with equivalent driving condition clusters of light vehicles. We can also see that the vehicles of clusters 2, 3 and 9 are always misclassified. This can be explained by the lack of data for both learning and testing in those clusters. We can finally state that heavy vehicles in acceleration (cluster 8) are confused half of the time with light vehicles in acceleration (cluster 5) and 25 % of the time in constant speed.

This *in-situ* classification test is promising and the performances seem coherent with the literature [15, 16, 28, 29]. This work confirms that the audio signal include the clues to identifies the vehicle types and driving conditions, such as humans can do.

### 3.3. Urban sound scene - Perceptual categories

A study carried out by Morel *et al.* [31] proposes a multi-criterion typology of road traffic pass-by noises. During a free clustering task, they found that the subjects were

Table IV. Morel *et al.* perceptual clusters.

| | 2-Wheeler | Light V. | Heavy V. |
|---|---|---|---|
| Const. Speed | 1 | 3 | 3 |
| Acceleration | 2 | 6 | 7 |
| Deceleration | 4 | 5 | 5 |

grouping pass-by noises into clusters that can mainly be explained using two criteria: the vehicle type and the vehicle driving condition. They were also interested in the influence of a third criterion, the road morphology, but it revealed not to be significant in the clustering process.

During the free clustering task, the subjects were gathering the light and heavy vehicles in both constant speed and deceleration. As a result, their proposed perceptual clustering of the road traffic is explained by seven clusters detailed in Table IV.

We propose here to modify the previous classification method to use Morel *et al.* perceptual clusters on the same training and testing sets than those used in Section 3.2. After a parametric study, the best success rate is found to be 84 % with the relaxation parameters

$$\mathbf{C} = [2.38, 1.68, 0.42, 0.9, 0.14, 0.14, 0.53].$$

The confusion matrix is detailed in Figure 9. It shows some good results, especially for clusters 1, 3 and 6: the classification rate for these clusters are equivalent to those obtained in Section 3.2 (previously labelled category 1, 4 and 5 respectively). We can see that the two-wheelers in deceleration (category 4) are still classified as a light (or heavy) vehicles in constant speed (category 3). We can also notice improvements: two-wheelers in acceleration are now well-classified 40 % of the time versus 0 % previously and the classification accuracy of heavy vehicles in acceleration (cluster 7) rises from 25 to 50 %. Nevertheless, classification rate decreased from 70 to 64 % for cluster 5 because it include the heavy vehicle decelerating that is still classified as passing-by in constant speed.

It is interesting to notice that the classification accuracy rises when we use those perceptual categories. This is not only because of the category merging (that can decrease the error), but also by improving the classification of some categories, such as the powered two-wheelers.

## 4. Monitoring over daytime

The classification method is applied to all the ten minute recordings acquired in one day at the Saint-Bernard Quay. Figure 10 shows the number of detected vehicles during the six measurements available with the cluster estimated by the method. The data presented in this Section are labelled by their starting acquisition time. The tracking method allowed to detect 539 vehicles during the acquisition time.

First, we can see that the number of detected vehicles decreases from midday to 15:00 then rises at 15:15. We

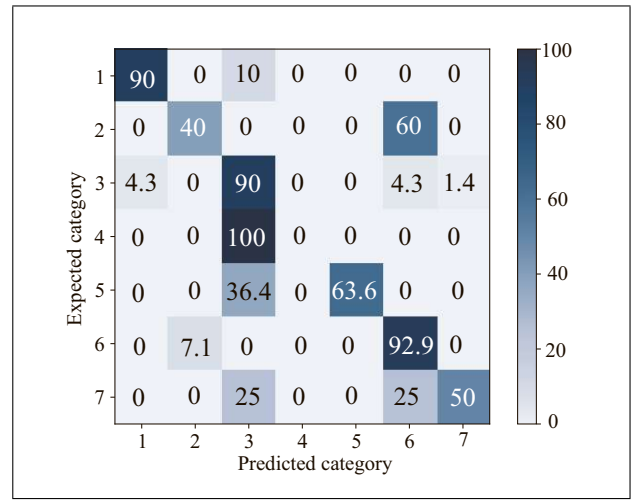


Figure 9. Confusion matrix in percentage of vehicle per perceptive cluster - Global classification accuracy: 84 %.

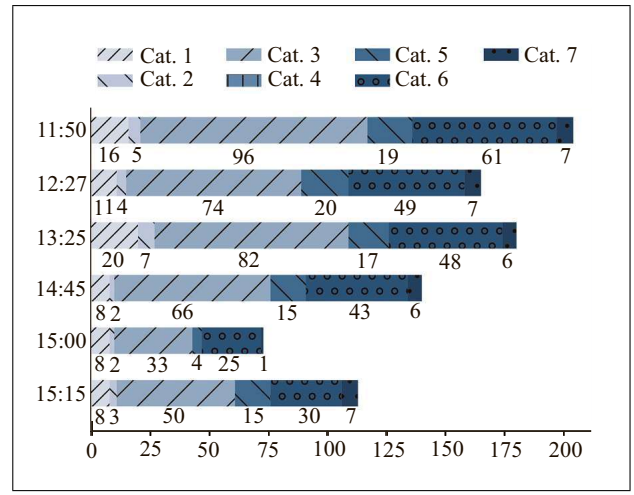


Figure 10. Number of vehicles over the different ten minute acquisitions by perceptual clusters. Total number of detected vehicles: 539.

can suppose that it continues rising later in the afternoon. Also, such as in Section 3.3, we can see that the 2-wheelers in deceleration (cluster 4) are never detected because of the camera position. We can also see that there is no major evolution of the distribution over the clusters, except at 15:00 when clusters 5 and 7 (light and heavy vehicles in deceleration) are less detected. We can finally notice that the road traffic is highly dominated by clusters 3 (light and heavy vehicles at constant speed) and 6 (light vehicles in acceleration) but not much vehicles in deceleration (categories 4 and 5). This can be explained by the street configuration (see Figure 3): three lanes located after a traffic light and only one before.

Figure 11 shows the evolution of the distribution of the equivalent sound level by perceptual cluster with boxplots. The sound level is calculated over the pass-by duration: one second (for the fastest vehicle) and five seconds for the accelerations and decelerations. Note that cluster 4 is never detected and not represented. We can notice a ten-

dency: the lower noise level values are emitted by the light and heavy vehicles decelerating (category 5). This is not the case for the 15:00 ten minute recordings. When analysing precisely the distribution it seems to be bi-modal with two sets of values centred on 61 dB and 78 dB. This bi-modality seems to be only due to the integration time and so the type of deceleration: short and stopping quickly (high equivalent level) or long and idling in front of the array (low equivalent level).

It seems also that the higher noise levels can be attributed to the 2-wheelers passing-by at constant speed (cluster 1) and heavy vehicles accelerating (cluster 7). For this last cluster, at 15:15, we have a large variation range due to two sets of values centred on 64 dB and 87 dB. It is interesting to note that this cluster has quite a constant number of vehicle during the day (between 6 and 7, except for the 15:00 experiment) but with a important variation of noise levels.

Moreover, we can notice that the variation range for each cluster is small until 13:25, reflecting an homogeneous traffic, but then, we have noticed large variations in the noise levels and an important number of outliers. This can reflect the increasing diversity of the road traffic for those periods (different types of heavy vehicles or different cylinders number for the 2-wheelers).

With these results, we can see that the analysis by perceptual category doesn't directly imply a reduced variability of noise level. Thus, we can see that both the traffic segmentation thanks to the classification task and the associated noise level are complementary in a urban noise monitoring system.

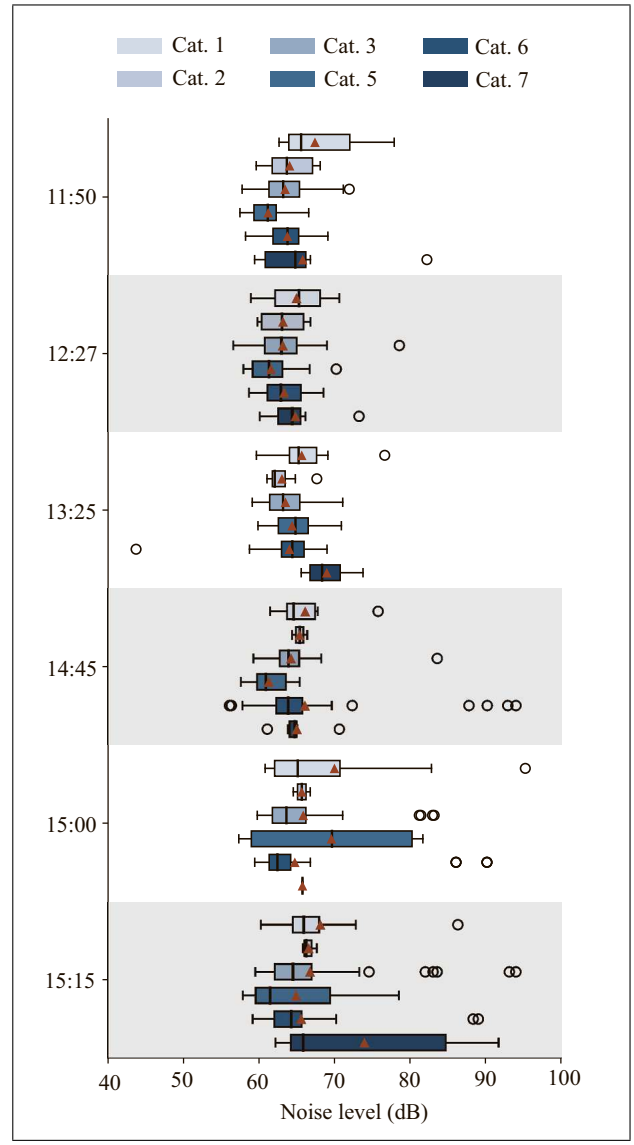


Figure 11. Boxplots of noise levels over the different ten minute acquisitions (labelled by their starting acquisition time) by perceptual clusters. Note that cluster 4 is never detected and not represented. The limits of the rectangles represent the first and third quartiles so that 50% of the data is included in this range. The red triangle symbolises the mean value, the black vertical bar the median value and the black circles the outliers.

## 5. Conclusion

In this study, we have been interested in proving the feasibility of monitoring the urban road traffic from the vehicle radiated sound field (thermal vehicles only). In order to do so, large arrays of microphones have been implemented to spatially filter a sound scene. This was achieved thanks to a dedicated beamforming algorithm coupled with a video tracking method, able to extract the audio signal of each passing-by vehicle. Appendix A1 presents the method validation with simulations and an isolated vehicle experiment. The *in situ* spatial filtering gain in also investigated.

Once the signal extracted, a classification step has been designed with Support Vector Machines using MFCCs as audio features. As learning samples, MFCCs where completed by the driving condition information based on the video tracking algorithm. It led to 99% of accurate classification on isolated vehicles.

Then, the application to a real urban sound scene has been presented with 539 detected vehicles. A maximum of 82% of accurate classification has been pointed out. This was mainly due to the lack of data for the 2- wheelers in acceleration and deceleration and for the heavy vehicles in deceleration. To reach this performance, the learning database is based on the isolated vehicle database but also on manually tagged pass-by in the urban *in situ* measurements. An adaptation is finally proposed to classify over perceptual clusters. It allowed to increase the accurate classification rate to 84%.

Finally, an application to six available ten minutes recordings has been presented. It allows to analyse the noise level for each perceptual category over the day. It has mainly pointed out that the most noisy vehicles – for this measurement place (St-Bernard Quay in Paris city centre) – where the 2-wheelers in constant speed and the heavy-vehicles in acceleration. The least noisy category was found to be almost always the light and heavy vehicles in deceleration.

This method appears to give a good knowledge of the road traffic composition. Nevertheless, it could be im-

proved by adding samples in the training and testing datasets, especially with two-wheelers and heavy vehicles decelerating signals but also with two-wheelers in acceleration. Note that the microphone array is easy to use and can be adapt in all urban situations (*eg.* on attaching it to the balconies). Even though the current one – in St-Bernard Quay – is very challenging, the results are already satisfying.

Subsequently, some improvements could be investigated. The video tracking step could be improved to rise the number of detected vehicles. It could be done using a remote camera with a better field of view or by coupling the video tracking with a tracking system on an acoustic image. In addition, in the presence of leaves, this method could be modified by doing the tracking step over an acoustic image rather than on the video. The performance should not be really degraded as the effect of the leaves should be at very high frequency.

The array geometry of the *in situ* experiment could also be improved in order to allow a better source separation between the traffic lanes. Furthermore, as we have extracted the audio signal of each vehicle, different kind of metric could be computed to better assess short-term noise annoyance in urban environments such as loudness or annoyance itself thanks to different models [32, 33]. There is no clear consensus in the literature on the link between audio signal and long-term annoyance (as used by WHO in the estimation of DALYs) but we assume that the short-term annoyance would probably explain a larger part of the variance of the long-term annoyance than any metric derived from the instantaneous or equivalent sound level.

## Appendix

### A1. Beamforming validations

We propose here some validation cases for the beamforming formalism on moving source. The propagation and beamforming model is first tested on simulated data. An adaptation is then proposed to extract audio signal without the energy compensation so that the output signal sound level is the one recorded by the microphones. This beamforming model is then validated on a similar experiment. Finally, extraction performance is investigated.

#### A1.1. Model accuracy

Based on the propagation and beamforming models presented in section 2.3, a simulation with a source moving at 50 km/h is proposed. It is done with a monopolar source emitting a 2 kHz pure tone with an amplitude of 1 Pa (90.9 dB SPL). The simulation lasts 3 s allowing the source to travel 41.6 m. Signals recorded by a linear microphone array in the configuration presented in Figure 4 (with $h = 16.5$ m) is simulated. The array is the same as the one used for Saint-Bernard Quay experiment (presented in Section 2.2.3): it is 21.6 meters long with 128 microphones regularly spaced 17 cm. The simulation is done at a sample rate of $F_s = 50$ kHz and computed in time domain rounding the reception time $t_r$ to the closest time sample of each microphone signal, no interpolation is introduced.

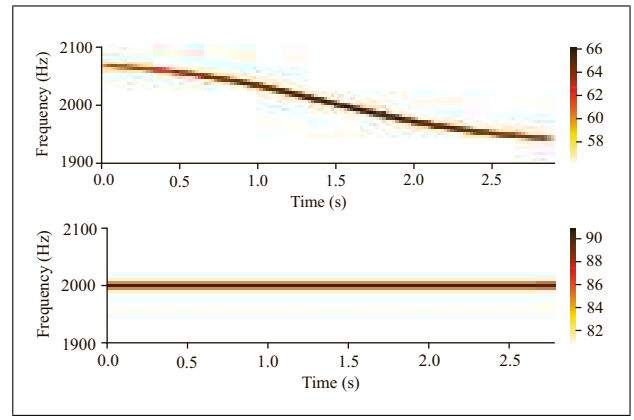


Figure A1. Spectrogram of simulated and beamformed signal for a source emitting a 2 kHz pure tone - $V = 50$ km/h - Dynamics = 10 dB. (a) Simulated signal, (b) Beamformed signal.

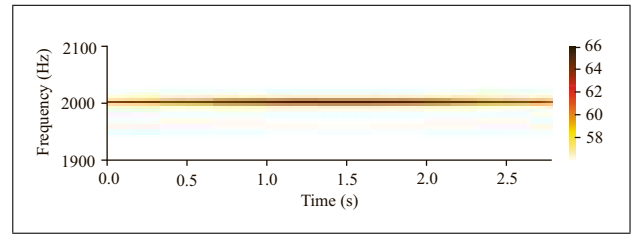


Figure A2. Spectrogram of beamformed signal realised with Equation (3), for a source emitting a 2 kHz pure tone - $V = 50$ km/h - Dynamics = 10 dB.

Figure A1a shows the spectrogram of the central microphone simulated signal where the Doppler effect is visible. Moreover, the amplitude is varying from 62 to 66 dB while the source approaches the microphone.

Figure A1b shows the spectrogram of the beamformed signal. We can see that the beamforming allows to get the initial emitted sound field, meaning that the amplitude remains constant at 90.9 dB with a frequency shift compensated.

This simulation shows a good accuracy of the beamforming to estimate the radiated sound field of a single moving source. However, we will be interested in removing the energy compensation so that the output signal sound level is the one recorded by the microphones. Indeed, by doing so, our results are more comparable to the noise maps (that provides $L_{den}$ in facade) and to city dwellers feeling. This can be done by removing the distance factor $r_{mi,e}$ in equation (2), so that the extracted signal is obtained in equation (3).

Figure A2 shows the spectrogram of the resulting beamformed signal for the simulation presented before. By comparing with the spectrogram of the initial signal (Figure A1a), we can see that the evolution of the energy is conserved (from 63 to 66 dB while the vehicle approaches).

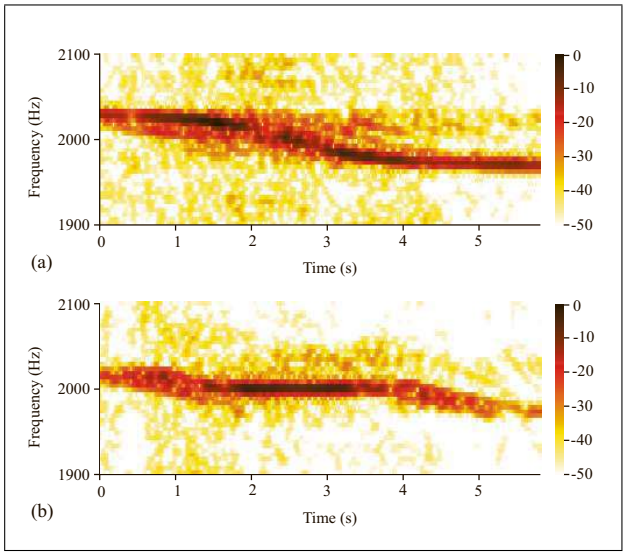


Figure A3. Spectrograms of recorded and beamformed signals for a source emitting a 2 kHz pure tone - $V = 20$ km/h - Dynamics = 50 dB. (a) Recorded signal, (b) Beamformed signal.

### A1.2. Validation on experimental data

The set-up and the beamforming method are proposed to be validated on a pass-by measure realised during the test-track experiment presented in Section 2.2.2. To do so, a loudspeaker emitting a 2 kHz pure tone has been set-up on a car passing-by at 20 km/h. The Figure 1a has shown the tracking step for this experiment. As we can see on this figure, the size of the moving object is over-estimated on the edge of the frame because it takes into account the shadows.

Figure A3 shows the spectrograms of the recorded and beamformed signals referenced to their maximum. On Figure A3a, we can notice both the frequency shift for the loudspeaker signal but also the broadband noise produced by the tire/road contact.

The beamformed signal spectrogram is presented in Figure A3a. The de-Dopplerization seems not to be perfect in the first and last seconds of the signal. This is due to the mis-positioning of the source when it enters and leaves the camera frame. Indeed, the beamforming is done on the centroid of the rectangle including the moving object so that we focus on the shadows and the engine when the car enters and on the exhaust pipe and the shadows when it leaves the frame. This involves a wrong speed estimation if the vehicle is not entirely in the camera frame. But when it is, we can see that the signal is well de-Dopplerised and the background noise is reduced proving the good estimation of the vehicle position with respect to the microphone array.

### A1.3. Extraction performances

The performances of this technique in filtering a sound scene is now investigated. For this purpose, during the *in situ* experiment (with the 128 microphone array) a loudspeaker was set-up at different places emitting a 1 kHz pure tone.

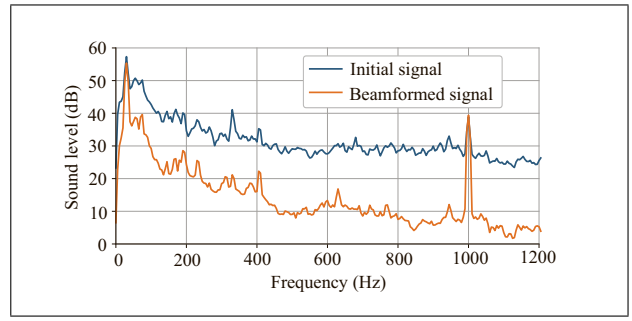


Figure A4. Power spectral density of central microphone (in blue) and beamformed signal on the 1 kHz pure tone loudspeaker. Distance between source and central microphone: 26.7 m, V = 0 km/h.

Figure A4 shows the power spectral density of the central microphone and the one of the beamformed signal while the source is placed at 26.7 m from the array centre. We can see that initially the loudspeaker is barely audible because of the energy of the other sound sources (road traffic). But thanks to the array dimension (21.6 m long) and beamforming we can see (orange curve) that the background noise is reduced by 20 dB around 1 kHz. But this gain reduces when the frequency decreases, reflecting the fact that the array resolution is frequency dependent. Such that in very low frequency (under 50 Hz) the technique seems not to provide any filtering gain.

## Acknowledgements

The authors wish to thank: Dominique BUSQUET (Sorbonne University), Jean-Christophe CHAMARD (PSA), Hélène MOINGEON (Sorbonne University), Christian OLLIVON (Sorbonne University) et Vincent ROUSSARIE (PSA).

This research benefits from the support of the chair "Mobilité et qualité de vie en milieu urbain" (Mobility and life quality in urban areas), carried by the Sorbonne University Foundation from 2014 to 2017 and supported by sponsors (PSA Peugeot-Citroën et Renault).

## References

[1] N. Misdariis, R. Marchiano, P. Susini, F. Ollivier, R. Leiba, J. Marchal: Mobility and life quality relationships – measurement and perception of noise in urban context. INTER-NOISE, Melbourne, 2014.

[2] W. Babisch: Traffic noise and cardiovascular disease: Epidemiological review and synthesis. Noise and Health **2** (2000) 9–32.

[3] D. Vienneau, C. Schindler, L. Perez, N. Probst-Hensch, M. Röösli: The relationship between transportation noise exposure and ischemic heart disease: A meta-analysis. Environmental Research **138** (apr 2015) 372–380.

[4] W. Babisch: Updated exposure-response relationship between road traffic noise and coronary heart diseases: a meta-analysis. Noise and Health **16** (2014) 1–9.

[5] L. Fritschi, A. L. Brown, R. Kim, D. Schwela, S. Kephalopoulos (eds.): Burden of disease from environmental noise. quantification of healthy life years lost in europe. WHO Regional Office for Europe, 2011.

[6] Directive 2002/49/EC relating to the assessment and management of environmental noise. European Parliament & Council of the European Union, 2002.

[7] B. Berglund, T. Lindvall, D. H. Schwela: Guidelines for community noise. 1999.

[8] Night noise guidelines for europe. World Health Organization, 2009.

[9] X. Sevillano, J. C. Socoró, F. Alías, P. Bellucci, L. Peruzzi, S. Radaelli, P. Coppi, L. Nencini, A. Cerniglia, A. Bisceglie, et al.: Dynamap–development of low cost sensors networks for real time noise mapping. Noise Mapping **3** (2016).

[10] J. Vos: Noise annoyance caused by mopeds and other traffic sources. INTER-NOISE, 2006.

[11] L.-A. Gille, C. Marquis-Favre, A. Klein: Noise annoyance due to urban road traffic with powered-two-wheelers: Quiet periods, order and number of vehicles. Acta Acustica united with Acustica **102** (may 2016) 474–487.

[12] Commission directive (eu) 2015/ 996 of 19 may 2015 establishing common noise assessment methods according to directive 2002/ 49/ ec of the european parliament and of the council.

[13] C. Marquis-Favre, E. Premat, D. Aubrée: Noise and its Effects – A Review on Qualitative Aspects of Sound. Part II: Noise and Annoyance. Acta Acustica united with Acustica **91** (2005).

[14] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze: Using one-class svms and wavelets for audio surveillance. IEEE Transactions on Information Forensics and Security **3** (Dec. 2008).

[15] X. Valero, F. Alias: Hierarchical classification of environmental noise sources considering the acoustic signature of vehicle pass-bys. Archives of Acoustics **37** (jan 2012).

[16] J. C. Socoró, F. Alías, 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** (2017).

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

[18] J.-R. Gloaguen, A. Can, M. Lagrange, J.-F. Petiot: Road traffic sound level estimation from realistic urban sound mixtures by non-negative matrix factorization. Applied Acoustics (2019) 229–238.

[19] E. Weinstein, K. Steele, A. Agarwal, J. Glass: Loud: A 1020-node microphone array and acoustic beamformer. ICSV14 Cairns • Australia 9-12 July, 2007.

[20] I. Hafizovic, C.-I. C. Nilsen, M. Kjølerbakken, V. Jahr: Design and implementation of a MEMS microphone array system for real-time speech acquisition. Applied Acoustics **73** (Feb 2012) 132–143.

[21] C. Vanwynsberghe, R. Marchiano, F. Ollivier, P. Challande, H. Moingeon, J. Marchal: Design and implementation of a multi-octave-band audio camera for realtime diagnosis. Applied Acoustics **89** (Mar 2015) 281–287.

[22] R. Leiba, F. Ollivier, R. Marchiano, N. Misdariis, J. Marchal: Urban acoustic imaging : from measurement to the soundscape perception evaluation. Inter-Noise, 2016.

[23] P. M. Morse, K. U. Ingard: Theoretical acoustics. McGraw-Hill, 1968.

[24] R. Cousson, Q. Leclère, M.-A. Pallas, M. Bérengier: Identification of acoustic moving sources using a time-domain method. Berlin Beamforming Conference, March 2018.

[25] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, M. D. Plumbley: Detection and classification of acoustic scenes and events: An ieee aasp challenge. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct 2013, 1–4.

[26] S. Davis, P. Mermelstein: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing **28** (August 1980) 357–366.

[27] C. Cortes, V. Vapnik: Support-vector networks. Machine Learning **20** (Sep 1995) 273–297.

[28] J. Salamon, J. P. Bello: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 2016, IEEE.

[29] K. J. Piczak: Environmental sound classificationwith convolutional neural networks. 2015 IEEE International Workshop on Machine Learning for Signal Processing, Boston, USA, Sep. 2015, IEEE.

[30] 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. J. Acoust. Soc. Am. **122** (2007) 881.

[31] J. Morel, C. Marquis-Favre, D. Dubois, M. Pierrette: Road traffic in urban areas: A perceptual and cognitive typology of pass-by noises. Acta Acustica united with Acustica **98** (Jan 2012) 166–178.

[32] J. Morel, C. Marquis-Favre, L.-A. Gille: Noise annoyance assessment of various urban road vehicle pass-by noises in isolation and combined with industrial noise: A laboratory study. Applied Acoustics **101** (Jan 2016) 47–57.

[33] A. Klein, C. Marquis-Favre, R. Weber, A. Trollé: Spectral and modulation indices for annoyance-relevant features of urban road single-vehicle pass-by noises. J. Acoust. Soc. Am. **137** (Mar 2015) 1238–1250.