EURASIP Journal on Audio, Speech and Music Processing manuscript No.
(will be inserted by the editor)

Estimation of the road traffic sound levels in urban areas based on non-negative matrix factorization techniques

Jean-Rémy Gloaguen · Mathieu Lagrange · Arnaud Can · Jean-François Petiot.

Received: date / Accepted: date

Abstract The advent of low cost acoustic monitoring devices raises new interesting approaches for improving the monitoring of the acoustic quality of urban areas. State of the art approaches target road traffic noise maps and consider, as input, an estimate of the number and the speed of vehicles in major traffic lanes. Follows a prediction procedure that outputs an acoustic pressure level at any location in the modeled area.

Considering as input the acoustic pressure measured in many locations using a sensor grid approach would greatly complement and improve the quality of the predicted pressure values. Among the technical issues that ~~raise~~ bring this kind of innovative approaches, there is a need to identify which part of the overall acoustic pressure level is due to the road traffic.

In this paper, several techniques based on non-negative matrix factorization framework are studied in this application scenario on a simulated sound scene corpus. The task ~~being~~ is to the best of our knowledge never been considered in the literature~~, we~~. We propose an experimental protocol to validate the studied approaches that complies with standard reproducible research recommendations. The results show the interest of our proposed approach for such sound environments as it improves the estimation of the road traffic sound level compared to basic methods.

Keywords non-negative matrix factorization · road traffic sound level · urban sound environment

J.-R. Gloaguen
Ifsttar Centre de Nantes UMRAE
Allée des Ponts et Chaussées 44344 Bouguenais
E-mail: jean-remy.gloaguen@ifsttar.fr

M. Lagrange
LS2N 1 rue de la Noe 44321 Nantes
E-mail: mathieu.lagrange@cnrs.fr

A.Can
E-mail: arnaud.can@ifsttar.fr

J.-F. Petiot
E-mail: jean-francois.petiot@ls2n.fr

1 Introduction

With the introduction of the European Directive 2002/-EC/49, cities over 100 000 inhabitants have to produce road traffic noise maps. These maps depict the sound level distribution over the city and an estimation of the number of city dwellers exposed to high noise levels. These maps play both an important communication role and help drawing up action plans to reduce noise exposure. Road traffic noise maps are the result of a simulation process based on the estimation of the traffic density on the main roads and the use of sound propagation modelling. They express as output L_{DEN} and L_N values, which are *Day-Evening-Night* and *Night* equivalent A-weighted sound levels respectively. ~~Altough~~ Although very useful, the produced noise maps introduce lot of uncertainty generated by the numerical tools [**?**] or by the different calculation methodologies used [**?**][**?**], despite the long data collection and calculation times. In addition, the usual road traffic noise maps are static, aggregating the exposure into indicators L_{DEN} and L_N, that ignore the sound levels evolution throughout the day. The use of acoustic measurements could facilitate their updating or even the generation of dynamic maps [**?**]. These measurements can be performed at fixed stations spread all over the cities [**?**] [**?**], which would make ~~available of~~ the long-term evolution of the traffic noise ~~levels~~available. It can also be performed with mobile stations [**?**] [**?**] covering

a larger area with fewer sensors but also sparse time periods.

[Fig. 1 about here.]

Currently, sensor networks in cities are spread for multiple applications (air quality assessment, measurement of meteorological parameters...), including the assessment of urban noise levels. The DYNAMAP project [?] studied the deployment and feasibility of such installations focusing on sensor installations on specific roads at the city scale in Milan and Rome [?]. The SONYC project (Sounds Of New-York City) aims to deploy a sensor network in New-York City for the purpose of monitoring constantly the noise pollution in the city [?]. In order to better know the urban sound environment, sensors are coupled with a detection tool that identifies the sound sources present [?]. In a similar way, but reduced to few neighborhoods with a denser network, the CENSE project[1] [?] aims to combine *in situ* observations, from a sensor network, and numerical data, from noise modeling, through data assimilation techniques.

Prior to data ~~assimiliation~~assimilation, the issue of the correct estimation of the traffic sound level from acoustic measurements ~~is still unsolved [?]. Mainly because~~ begin to be studied [?,?]. As the urban sound environment is a complex environment gathering lots of different sounds (car passages, voices, whistling bird, car horn, airplanes...) that overlap~~. Consequently~~, the traffic sound level estimation based on measurements is not a trivial task.

2 Related works

Many recent works have focused on the detection or recognition tasks of environmental sounds [?], [?], [?], [?]. A two-step process is generally followed: describe the audio files with a set of features (Spectral Centroid, harmonicity, Mel-Frequency Cepstral Coefficient ...) and classify them with the help of classifiers (Support Vector Machines, Gaussian Mixture Models, Hidden Markov Model, Artifical Neural Networks). A description of these features and classifiers can be found in [?] and their applications can be found in [?], [?], [?].

The main issue in the detection or recognition tasks is the overlap of environmental sounds. Although near major roads, traffic is predominant, there are many places where it overlaps with other sound sources which contribute significantly to the overall sound levels. To circumvent this issue, Socoró et al. propose to suppress time frames where there is significant overlap by considering an Anomalous Noise Events Detector [?]. It consists in detecting the unwanted sound sources from labeled recordings, *i.e.* that are not related to the traffic component. Those time frames are then discarded in order not to take them into account during the estimation of the traffic sound level. An alternative approach that we will follow in this paper is to consider the blind source separation paradigm to reliably estimate the traffic noise level, see Figure 1. It consists in separating the contribution of the traffic from the other sources within a polyphonic scene. One major advantage of following such approach is that the estimate is continuously available, making the approach applicable in a wide range of urban areas, even where the traffic noise is relatively low compared to the remaining contributions.

In an urban environment context, source separation can be achieved with the help of acoustic microphone arrays and beamforming [?]. However, this approach requires spreading multiple microphones arrays in cities that is very expensive (even with low cost microphones) and time-consuming for calibration and maintenance. This method is then not considered here to be deployed all over cities. ~~In the opposite~~On the contrary, monophonic sensor networks need less microphones but the main challenge is to succeed to estimate correctly the road traffic from only one signal in which all kind of sound sources can be present. A convenient method for this is the Non-negative Matrix Factorization (NMF) technique [?]. When considering audio as input, it usually consists in approximating the magnitude spectrogram of an audio file by the product of two low rank matrices, one representing the components of interest and the other the contribution at a given time of those components to approximate the input magnitude spectrogram [?] [?]~~[?]~~. In the audio processing domain, NMF has already been employed for the source separation task of monaural signals of speech and music [?] [?]. By design, this method deals reasonably well with the overlapping sound sources as soon as the overlap can be resolved on the time/frequency plane.

Closer to our application scenario, NMF has been considered ~~by~~ in [?] where coupled NMF has been used as a sound event detection in real life recordings. However, no information on the quality of the detected signal is mentioned. Nevertheless, they raised the question of the dictionary size reduction after the learning phase: cluster the elements and keep the full spectrum or use mel spectrum representation and keep the full learnt dictionary? They conclude on the efficiency of each approach

[1] `http://cense.ifsttar.fr/`

and suggere to combine both to deal with large databases. Also, Innami and Kasai [?] ~~. After having performed~~ proposed to perform NMF on simulated audio files ~~, they perform~~ as a source separation ~~in~~ method. Their tool is composed of two steps by 1) separating the sound background from the events and 2) by isolating the events using spectral features using a k-means procedure. ~~We study in this paper different flavors of~~ On a preliminary study [?], supervised Non-Negative Matrix Factorization ~~where the traffic component is consideredin its entirely whether it is a sound background or an event. We demonstrate that supervised NMF and semi supervised NMF approaches have some interests but fail to give satisfactory results for the application at hand. We thus introduce another NMF scheme called thresholded initialized NMF that makes good use of prior knowledge about the source of interest, in our case the traffic noise, but also generalizes well to several kinds of urban areas and to traffic to interference ratio (TIR)~~ has been considered, on simple simulated sound scenes, as a source separation method to extract the road traffic component (which include continuous traffic noise and passing car) in order to estimate its sound level.

~~To perform the numerical experiments, we consider a~~ Here, we extend this approach with more flavors of NMF on more complex sound scenes. This corpus of simulated sound scenes is created from a built-up sound database composed of a high number of diverse sound samples in order to encompass a wide variety of sound environment with a variable presence of road traffic noise. The use of simulated sound scenes ~~is mandatory for~~ allows rigorous experimental validation as it offers a high level of control on the design of the scenes and the knowledge of the exact contribution of the traffic component ($L_{p,traffic}$), which would be difficult to extract from a recording of an urban scene. The aim is to find the form of NMF that gives the minimal error on the reconstruction of the traffic noise signal on the whole corpus. We demonstrate that supervised NMF and semi supervised NMF approaches have some interests but fail to give satisfactory results for the application at hand. We thus introduce another scheme called thresholded initialized NMF that makes good use of prior knowledge about the source of interest, in our case the traffic noise, but also generalizes well to several kinds of urban areas and to traffic to interference ratio (TIR).

The remaining of the paper is organized as follows. Section 3 details the technical aspects of NMF and describes the 3 approaches considered in this paper to achieve the task at hand. Section 4 describes the corpus of environmental sound scenes and the experimental protocol setup. Section 6 presents and discusses the outcomes of the numerical results.

3 Non-negative Matrix Factorization

3.1 Description of NMF

Non-negative Matrix Factorization is a linear approximation method introduced by Lee and Seung, [?], which can be used to approximate the spectrogram $\tilde{\mathbf{V}}$ (obtained using a Short-Term Fourier Transform) of an audio file, $\mathbf{V}$, $\in \mathbb{R}^{+}_{F\times N}$ as :

$$\mathbf{V} \approx \tilde{\mathbf{V}} = \mathbf{WH} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{+}_{F\times K}$ is the *dictionary* (or basis) matrix composed of audio spectra and $\mathbf{H} \in \mathbb{R}^{+}_{K\times N}$ is the *activation* matrix which summarizes the temporal evolution of each element of $\mathbf{W}$. As the constraint of non-negativity of $\mathbf{W}$ and $\mathbf{H}$ is considered, NMF allows only additive combinations between the element of $\mathbf{W}$. It is then a part-based representation that NMF provides. An illustrative example can be found in Figure 2.

[Fig. 2 about here.]

The choice of the dimensions is often made so that $F\times K + K\times N < F\times N$ [?]. NMF is then considered as a low rank approximation method. However, this constraint is not mandatory. To estimate the quality of the approximation, an objective function is used

$$\min_{\mathbf{H}\geq 0,\mathbf{W}\geq 0} D\left(\mathbf{V}||\tilde{\mathbf{V}}\right). \tag{2}$$

The operator $D(x|y)$ is a divergence calculation such as:

$$D\left(\mathbf{V}||\tilde{\mathbf{V}}\right) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_\beta\left(\mathbf{V}_{fn}|[\mathbf{WH}]_{fn}\right) \tag{3}$$

and usually belongs to the $\beta-$divergence class [?] in which the well known Euclidean distance (eq. 4a) and the Kullback-Leibler divergence (eq. 4b) belong

$$d_\beta(x|y) = \begin{cases} \frac{1}{2}(x-y)^2, & \beta = 2, \quad (4a) \\ x\log\frac{x}{y} - x + y, & \beta = 1. \quad (4b) \end{cases}$$

To better take into account prior knowledge about the sources of interest, constraints (like the smoothness

or the sparsness criteria [?]) can be added to the objective function.

Algorithms have been proposed to solve the minimization problem (2) iteratively such as the multiplicative update [?], the alternating least square method [?] or the projected gradient [?]. Here, the multiplicative update is chosen as it has been well studied in the literature and it ensures convergence of the results [?].

3.2 Supervised NMF

First, supervised NMF (Sup-NMF) is used: the *dictionary* includes audio spectra of urban sound sources. A lot of the different sound sources present in the urban environment are known. Their spectra can be obtained and be a basis of $\mathbf{W}$. The *activation* matrix is then the unknown variable to estimate. In the first iteration, $\mathbf{H}$ is initialized randomly, then it is updated by the generic algorithm [?]

$$\mathbf{H}^{(i+1)} \leftarrow \mathbf{H}^{(i)} . \left(\frac{\mathbf{W}^T \left[\left(\mathbf{W}\mathbf{H}^{(i)} \right)^{(\beta-2)} . \mathbf{V} \right]}{\mathbf{W}^T \left[\mathbf{W}\mathbf{H}^{(i)} \right]^{(\beta-1)}} \right)^{\gamma(\beta)} \tag{5}$$

with $\gamma(\beta) = \frac{1}{2-\beta}$, for $\beta < 1$, $\gamma(\beta) = 1$, for $\beta \in [1,2]$ and $\gamma(\beta) = \frac{1}{\beta-1}$ for $\beta > 2$. The product $A.B$ and A/B are respectively the Hadamard product and ratio. As in the supervised approach the indexes of traffic components in $\mathbf{W}$ are known, the separation of the corresponding sound source is made by extracting the related basis and activators,

$$\tilde{\mathbf{V}}_{traffic} = [\mathbf{W}\mathbf{H}]_{traffic} . \tag{6}$$

3.3 Semi-supervised NMF

The supervised approach is useful when prior knowledge can be assumed for all the sources in the mixture, which is not a reasonable assumption in our application scenario. To some ~~extend~~extent, prior knowledge can be considered for the traffic but not for the numerous kind of interferences that can occur in a realistic scenario. To better take into account the diverse nature of urban scenes, semi-supervised NMF (Sem-NMF)[?] is a good candidate as it offers more flexibility. This method assumes a *dictionary* with a fixed part $\mathbf{W_s} \in \mathbb{R}^+_{F\times K}$, composed in our case of road traffic spectra, and with a mobile part, $\mathbf{W_r} \in \mathbb{R}^+_{F\times J}$ with $J << K$, that is updated during optimization. In the literature, J is set to a small number with respect to K so as to force the optimization to still consider the fixed part of the dictionary [?]. The aim is to include in $\mathbf{W_r}$ the elements that are not related with the traffic. The problem (1) becomes

$$\mathbf{V} \approx \tilde{\mathbf{V}} = \mathbf{W_s}\mathbf{H_s} + \mathbf{W_r}\mathbf{H_r} \tag{7}$$

with $\mathbf{W} = [\mathbf{W_s}\mathbf{W_r}]$ and $\mathbf{H} = \begin{bmatrix} \mathbf{H_s} \\ \mathbf{H_r} \end{bmatrix}$. In a similar way as to solve the equation (2), $\mathbf{W_r}$, $\mathbf{H_r}$ and $\mathbf{H_s}$ are successively updated with the relations (8):

$$\mathbf{W_r}^{(i+1)} \leftarrow \mathbf{W_r}^{(i)} . \left(\frac{\left[\left(\mathbf{W_r}\mathbf{H_r}^{(i)} \right)^{(\beta-2)} . \mathbf{V} \right] \mathbf{H_r}^T}{\left(\mathbf{W_r}\mathbf{H_r}^{(i)} \right)^{(\beta-1)} \mathbf{H_r}^T} \right)^{\gamma(\beta)} , \tag{8a}$$

$$\mathbf{H_r}^{(i+1)} \leftarrow \mathbf{H_r}^{(i)} . \left(\frac{\mathbf{W_r}^T \left[\left(\mathbf{W_r}\mathbf{H_r}^{(i)} \right)^{(\beta-2)} . \mathbf{V} \right]}{\mathbf{W_r}^T \left(\mathbf{W_r}\mathbf{H_r}^{(i)} \right)^{(\beta-1)}} \right)^{\gamma(\beta)} , \tag{8b}$$

$$\mathbf{H_s}^{(i+1)} \leftarrow \mathbf{H_s}^{(i)} . \left(\frac{\mathbf{W_s}^T \left[\left(\mathbf{W_s}\mathbf{H_s}^{(i)} \right)^{(\beta-2)} . \mathbf{V} \right]}{\mathbf{W_s}^T \left(\mathbf{W_s}\mathbf{H_s}^{(i)} \right)^{(\beta-1)}} \right)^{\gamma(\beta)} . \tag{8c}$$

Applications of Sem-NMF for speech denoising from background noise or musical content can be found in [?] and [?].

3.4 Thresholded initialized NMF

As it will be demonstrated in the experimental results described in Section 6, those last approaches fail to provide consistent results in a wide range of urban areas and for different traffic preponderances due to generalization capabilities issues.

We therefore propose an alternative scheme based on the unsupervised NMF. Usually in unsupervised learning, $\mathbf{W}$, as $\mathbf{H}$, is initialized randomly. Here, as the concerned sound source is known and audio samples of car passages are available, an initial dictionary, $\mathbf{W_0}$, is learnt by converting the audio files in the spectra domain; see part 4.2.1. Then NMF is performed where $\mathbf{W}$ (eq. 9) and $\mathbf{H}$ (eq. 5) are updated alternatively. $\mathbf{W}$ is therefore updated by forcing its ~~initiation~~ initialization with *a priori* knowledge but allowing it to adapt to the actual content of the scene under study,

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

After N iterations, a measure of similarity $D_\theta(\mathbf{W_0}||\mathbf{W})$ between $\mathbf{W_0}$ and the obtained dictionary $\mathbf{W}$ for each element k is computed using a cosine similarity metric,

$$D_\theta(\mathbf{W_0}_k||\mathbf{W}_k) = \frac{\mathbf{W_0}_k.\mathbf{W}_k}{||\mathbf{W_0}_k||.||\mathbf{W}_k||}. \quad (10)$$

$D_\theta(\mathbf{W_0}_k||\mathbf{W}_k) = 1$ means that the k-th element of $\mathbf{W}$ is identical to the k-th element of $\mathbf{W_0}$. ~~In the opposite~~On the contrary, $D_\theta(\mathbf{W_0}_k||\mathbf{W}_k) = 0$ means that the elements are completely different. This measure has the advantage to be bounded between 1 and 0 and to be invariant with respect to scale. The ~~k~~ K values of $D_\theta(\mathbf{W_0}||\mathbf{W})$ are next sorted in ~~descending~~ decreasing order. The elements in $\mathbf{W}$ that can belong to $\mathbf{W}_{traffic}$ are selected by a *hard thresholding* method. It is defined as:

$$\mathbf{W}_k \in \mathbf{W}_{k,traffic} \quad \text{iff} \quad D(\mathbf{W_0}_k||\mathbf{W}_k) > t \quad (11)$$

where t is a fixed threshold $\in$ [0;1]. An illustrative example is displayed in Figure 3.

[Fig. 3 about here.]

This approach is named *Thresholded inititialized NMF* (TI-NMF). Other thresholding methods as the *soft* [?] and the *firm* [?] have been investigated. A ~~fast~~ quick parametric study revealed that the *hard* thresholding method, as presented in Figure 3, was the most reliable approach.

4 Experimental protocol

In order to validate the usefulness of the proposed NMF scheme to estimate the road traffic noise levels, one need to compare the traffic noise level predicted by the algorithm to a reference level. The latter can hardly be measured or even annotated from real life recordings. Thus, we propose to consider simulated sound scenes to assess the performance of the proposed NMF. This offers a controlled framework to design at low cost a wide diversity of sound environments in which all the traffic components are known, thus allowing the computation of the reference level.

4.1 Environmental sound scene corpus

The corpus is designed with the *SimScene* software[2]. *SimScene* [?] is a simulator that creates monaural sound scenes in a .wav format by sequencing and summing audio samples that come from an isolated sound database. The simulator has been succesfully considered for a wide range of experimental design for sound detection algorithm assessment [?] [?] [?].

This database is divided into two categories: *i*) the *event* category, which are the brief sounds (from 1 to 20 seconds) that are considered as salient including 245 sound event samples divided in 19 sound classes (*ringing bell, whistling bird, sweeping broom, car horn, passing car, hammer, drill, coughing, barking dog, rolling suitcase, closing door, plane, siren, footstep, storm, street noise, metallic noise, train, tramway, truck and voice*) and *ii*) the *background* or *texture* category that includes all the sounds that are of long duration and whose acoustic properties do not vary with respect to time. 154 sound samples that belong to this category are divided in 9 sound classes (*whistling bird, construction site noise, crowd noise, park, rain, children playing in schoolyard, constant traffic noise, ventilation, wind*). These sounds are in .wav format sampled at 44.1 kHz. The sound class *passing car* comes from 60 recordings of 2 cars (Renault Megane and Renault Scenic) made on the Ifsttar's runway at different speeds with multiple gear ratios. The other audio files have been found online (*freesound.org*) and within the *salamon2014dataset* database [?]. Each sound class is composed of multiples samples (*bird01.wav*, *bird02.wav* ...) to allow some diversity in the resulting mixture, see Figure 4. The software allows the user to control some high level parameters (number of events of each class that appear in the mixture, elapsed time between each sample of a same class, presence of a fade in and a fade out ...) completed with a standard deviation that may bring some random behavior between the scenes. Furthermore, an audio file of each sound class present in the scene can be generated that allows us to know its exact contribution as well as a text file that summarizes the time presence of all the events.

[Fig. 4 about here.]

This database allows the creation of a wide diversity of ~~realistic~~ urban sound scenes from the road traffic point of view [?]. A sound mixing corpus is composed of 6 subcorpus of 25 audio files each lasting 30 seconds. Each sub-corpus is characterized by a specific generic sound class that summed with traffic will make the estimation of the traffic level more difficult. The classes are: *alert* (car horn, siren), *animals* (barking dog , whistling bird), *climate* (wind, rain), *humans* (crowd noise and voice), *mechanics* (metallic and construction site noises) and

[2] Open-source project available at: `https://bitbucket.org/mlagrange/simScene`

transportation (train, tramway and plane). In each file, the traffic component is present~~as~~. What is called traffic component is the sum of the ~~background and event traffic sounds while~~ road traffic background noise and the sound events generated by the passing car class. On the contrary, the *interfering* sound class ~~is the~~ includes all the other sound sources not related to it. Car horn sound class belongs to this component as it is considered as a warning signal. To test different scenarios, each audio file is duplicated with the traffic sound level of the entire sound scene, $L_{p,traffic}$, fixed to a specific level according to the sound level of the interfering class, $L_{p,interfering}$, following the relation (12).

$$TIR = L_{p,traffic} - L_{p,interfering} \quad (12)$$

with TIR , the *Traffic Interference Ratio* ~~$TIR \in$ {-12, -6, 0, 6, 12} dB~~. When $TIR < 0$ dB, the traffic component is less present than the interfering class. ~~In the opposite~~On the contrary, for $TIR > 0$ dB, the traffic class is louder than the interfering class. In most of the urban sound environments, the ratio between the interfering class and the traffic is mainly included between TIR = -6 dB and TIR = 6 dB [?]. This is between these values that the estimator has to be the most efficient. But, in this experimental framework, this frame is extended to TIR = -12 dB and TIR = 12 dB to test the limit of NMF. The total number of scenes designed is then 750 (6 sub-corpus × 25 scenes × 5 TIR values), each scene during 30 seconds, it leads to a full duration of 6 hours and 30 minutes.

4.2 Experiment

The experiment consists in estimating the road traffic sound level of the 6 environmental sound sub-corpus (*alert* (al.), *animals* (an.), *humans* (hu.), *climate* (cl~~),~~ .), ~~*mechanics*~~*transportation* (~~me),~~ tr.), ~~*transportation*~~*mechanics* (~~tr~~me.)) composed each of 25 audio files ($M = 25$) and for 5 TIR ({-12, -6, 0, 6, 12} dB). The range of these values is large but in the urban environments, the TIR seems to be between -6 dB and 12 dB [?]. The case TIR = -12 dB is then an extreme case to study the NMF behavior. The spectrogram $\mathbf{V}$ of each sound scene is built with a window size $w = 2^{12}$ with a 50 % overlap, see Figure 5.

[Fig. 5 about here.]

Assuming that the traffic spectral profile is largely concentrated in the low frequency components, a first estimator to determine the traffic sound level is a frequency low-pass filter. It depends only on the cut-off frequencies $f_c \in \{500, 1k, 2k, 5k, 10k, 20k\}$ Hz. The spectrogram $\mathbf{V}$ is filtered and the remaining energy is then considered as traffic component (eq. 13),

$$\tilde{\mathbf{V}}_{traffic} = \mathbf{V}_{f_c}. \quad (13)$$

The second estimator is the proposed scheme, based on the three NMF approaches presented in Section 3. Multiples experimental factors are involved here between the dictionary learning and NMF (see Figure 6), each experimental factor having multiples modalities.

[Fig. 6 about here.]

[Table 1 about here.]

4.2.1 NMF Dictionary

In order to prevent potential overfitting issues, the dictionary is built from a separate sound database dedicated specifically to this task. ~~It~~ The train database is composed of 53 audio files of ~~passing cars~~isolated passing cars with a 18 minutes and 29 seconds cumulative duration. These recordings have been made on the Ifsttar's runway too, with the same experimental conditions that the recordings of the *SimScene* database but with two different cars (Dacia Sandero and Renault Clio). The different steps leading to a dictionary is resumed in Figure 7. First, for each audio file, its spectrogram is calculated with fixed parameters (w, 50 % overlap, $nfft$). Then time/frequency windows of $F \times w_t$ dimensions are applied without overlapping on the spectrogram in order to consider several spectra for each audio file where $w_t \in \{0.5, 1\}$ second. In each window, the root mean square value is calculated on each frequency bin to reduce the windowed spectrogram in one spectrum a of $F \times 1$ dimension. With this size of window, it is possible to obtain the characteristic pitches of the different audio samples. One obtains for each value of w_t, from the 53 audio samples of passing cars respectively 2218 and 1109 elements. Since the number of elements given by this processing is high, in order to reduce the computational time and avoid redundant information, a K-means clustering algorithm is applied to reduce the number of spectra to $K \in \{25, 50, 100, 200\}$. The K centroids are then the elements considered in the dictionary. A special case is added where the root mean square of *all* the spectrogram is applied ($w_t = all$) to build a dictionary with the spectral envelope of each audio sample. In this case, 53 spectra are obtained. The K-means clustering algorithm is then reduced to $K \in \{25, 50\}$.

[Fig. 7 about here.]

[Fig. 8 about here.]

An example that illustrates the process can be found on Figure 8 on a 3 second extract of the spectrogram of a car passage, see Figure 8a. In the case where $w_t = 1$ second, 3 elements are therefore extracted from the spectrogram while in the case where $w_t = all$, all the spectrogram is reduced into one element, see Figure 8b.

~~The obtained dictionary is expressed with third octave bands and each~~ Each basis vector of $\mathbf{W}$ is normalized such as $||\mathbf{W_k}|| = 1$ with $||\bullet||$ the ℓ-1 norm. Table 1 summarizes the experimental factors (K and w_t) for the dictionary building and their related modalities. ~~This last step allows us to reduce the dimensionality while preserving a rich description of the spectral content. Experimental validation consistently showed that considering octave bands do not impact the performance of the estimator studied in this paper~~The 10 versions of the built dictionary are then used for NMF. In Sup-NMF, these 10 versions correspond to $\mathbf{W}$, in Sem-NMF, they correspond to the fixed part $\mathbf{W_s}$ and for TI-NMF, they are the initial dictionaries $\mathbf{W_0}$ that are next updated.

[Table 2 about here.]

[Fig. 9 about here.]

In the case of Sup-NMF, the MAE errors are important for all sub-classes ~~. This method~~ at low TIR. This approach reveals to be too rigid as $\mathbf{W}$ is composed of fixed traffic spectra. In the ~~case of low TIR, in the~~ aim to reduce the objective function, see eq. 2, traffic elements are used whatever the sound event in the sound scene. Thus forcing the dictionary to be only composed of traffic spectra is not a sufficient way to estimate correctly the traffic sound level, $\tilde{L}_{p,traffic}$. At high TIR, this approach generates better estimations as it is a favorable case: the *traffic* is the main component with a dictionary dedicated to this sound source.

~~Consequently, with~~ With the addition of a mobile part in the dictionary, $\mathbf{W_r}$, the semi-supervised approach allows a better consideration of the interfering class in low TIR. It brings a significant decrease of the errors for low TIR. However, the relatively high degrees of freedom of Sem-NMF are restrictive for high TIR as the errors exceed 2 dB for all sub-classes and increase with $TIR \in \{6, 12\}$ dB. The performances are even superior to the LP filter baseline. In order to reduce $D(\mathbf{V}||\mathbf{WH})$, without constraint, Sem-NMF is free to include traffic components in $\mathbf{W_r}$. Consequently, this behavior decreases the quality of the reconstruction of the traffic component.

~~Finally,~~ The TI-NMF behavior, with a threshold ~~$t = 0.42$~~ $t = 0.41$ and $\beta = 1$, ~~offers the lowest average error (Table 2)~~generates error inferior to the baseline for all the TIR values, with the exception at $TIR = 0$ dB. Its behavior is singular because it does not propose on each TIR the lowest error even if it is the best approach on all the corpus. Unlike Sup-NMF, where $\mathbf{W}$ is fixed, and Sem-NMF, where only $\mathbf{W_r}$ is updated, TI-NMF updates $\mathbf{W}$ entirely to adjust prior knowledge to the scene under evaluation so as to adapt to the different sound environments. The closest elements of the traffic component defined in $\mathbf{W_0}$ are then extracted to deduce the traffic signal.

For low TIR, as the traffic sound class is not predominant, the final dictionary $\mathbf{W'}$ strongly differs from ~~the~~ $\mathbf{W_0}$. With the thresholding, only a reduced number of basis vectors are considered as traffic components: from the 200 basis contained in $\mathbf{W'}$, 106 are considered as *traffic* component in TIR = -12 dB. In comparison to supervised results, this approach reduces significantly the error for the *human* and *transport* subclasses. However, for *climate* and *mechanics*, the errors remain important as these interfering classes have similar spectral profiles when compared to traffic ones.

For high TIR, as the traffic is the main sound source, the similarity of the initial dictionary and $\mathbf{W}$ is higher which allows retaining more elements as traffic components (181 basis for TIR = 12 dB) and then decreases the error ($MAE < 1$~~dB~~). The kept elements are then more suited to the scenes than a fixed dictionary. The error for these TIR is then due to the thresholding which put aside some elements that ~~are nevertheless~~ can be related to the traffic component. ~~In Figure 10, the 1 second equivalent sound pressure level of an alert scene is displayed for $TIR \in \{-12, 12\}$ dB for the 500 Hz low pass filter and for TI-NMF. In the case of $TIR = -12$ dB, the traffic elements selected in TI-NMF do not activated when the alarm class sounds contrary to the low-pass filter.~~

~~With a low threshold, it is~~ It would be possible to decrease the error ~~. For~~ by adapting the threshold according to the TIR for instance: at TIR = 12 dB, the average error on all the sub-classes ~~is 0.22~~ would decrease to 0.30 (± ~~0.08) dB~~ 0.11) with t = ~~0.30. In opposite for~~ 0.34 where 191 elements are considered as traffic component. In the opposite, at TIR = -12 dB, ~~it is better to choose a high threshold $t = 0.55$~~by increasing the threshold to 0.53, the error ~~decrease then to 4.46~~ would decrease to 4.49 (± ~~1.66) dB. In~~ 1.75) with in average 73 elements considered as *traffic* elements.

But, in order to generalize this method to a practical case where no prior knowledge on the urban environ-

ment is made, the chosen threshold t is then fixed to $t = 0.42$ 0.41 as it is the one that best balanced these opposite cases. In Figure 10, the 1 second equivalent sound pressure level of an alert scene is displayed for $TIR \in \{-12, 12\}$ dB for the 500 Hz low pass filter and for TI-NMF. In the case of $TIR = -12$ dB, the traffic elements selected in TI-NMF are not activated when the alarm class sounds contrary to the low-pass filter.

[Fig. 10 about here.]

5 Conclusion

In this work the non negative metric factorization framework was matrix factorization framework is used to estimate the road traffic sound level in urban sound mixtures. It is a well suited approach to these sound environments because it easily takes into account the overlap between the multiple sound sources present in the cities and it is adapted to monophonic sensor networks. Different versions of NMF have been studied as a supervised and semi-supervised approach. On a large corpus of sounds, the supervised approach proves to be too restrictive to be adapted to different sound environments whereas the semi-supervised approach has, on the contrary, too many degrees of freedom on the mobile dictionary $\mathbf{W_r}$, decreasing its performance especially when the traffic is predominant. The proposed approach, named threshold initialized NMF achieves the lowest average error on the entire corpus. With this method, the $\mathbf{W}$ is initialized with road traffic spectra, updated and the dictionary elements that are similar to the road traffic spectra are then extracted by hard thresholding. The study of the error according each TIR reveals that this method is not the most efficient on each TIR , but it has to be remind that, this parameter is not available in practical case. TI-NMF can be considered as the approach that limit the error in the different environment. A major advantage of the proposed approach is that it is not designed for a specific source. Even It has to be noticed that the method would gain to be tested on larger dataset with, for instance, motorcycles, noisy traffic elements which are absent here. However, even though the experiments described in this paper focused on road traffic sounds, changing the dictionary to contain bird sounds would lead to an estimator of the presence of birds. Extending the apporach to other sources is thus of intetrest interest for future research. Also, performance improvement could be achieved by the addition of constraints such as sparsness [?] and smoothness [?] of the low rank matrices.

The experimental protocol and the evaluated estimators have been implemented with the Matlab software. For reproducible purposes, the code is available online. The evaluation database composed of multiple samples of urban sounds is also made available for the research community with interest in detection, separation and recognition tasks of urban sound sources.

Declarations

Availability of data and materials

The dataset supporting the conclusions of this article is available in the *urban_traffic_nmf_dataset.zip* repository, `https://zenodo.org/record/1145855#.Wl2oPnkiGos`. The software supporting the conclusions of this article is available in

- Project name: article2017EstimationAmbiance
- Project home page: `https://github.com/jean-remyGloaguen/article2017EstimationAmbiance`
- Archived version: `https://doi.org/10.5281/zenodo.1145855`
- Programming language: Matlab
- Other requirements: Matlab 2016b or higher
- License: GNU GPL
- Any restrictions to use by non-academics: license needed

Competing interests

The authors declare that they have no competing interests.

Funding

This study is co-funded by Ifsttar and Pays de la Loire region with a partial funding from the ANR under project reference ANR-16-CE22-0012.

Authors' contributions

JG carried out the numerical experience experiment and drafted the manuscript. ML, AC and JP participated in the design of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

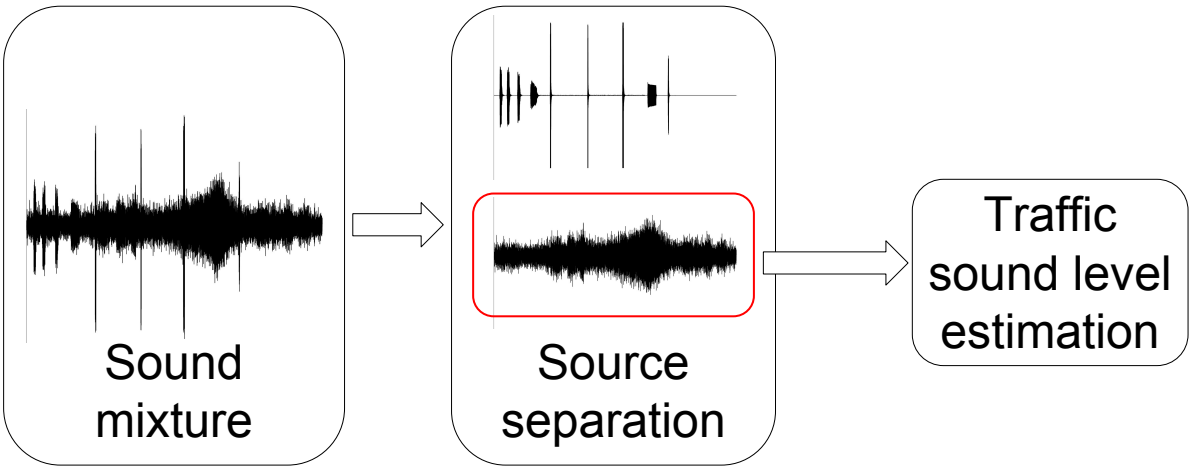


Fig. 1: Block diagram of the general source separation model.

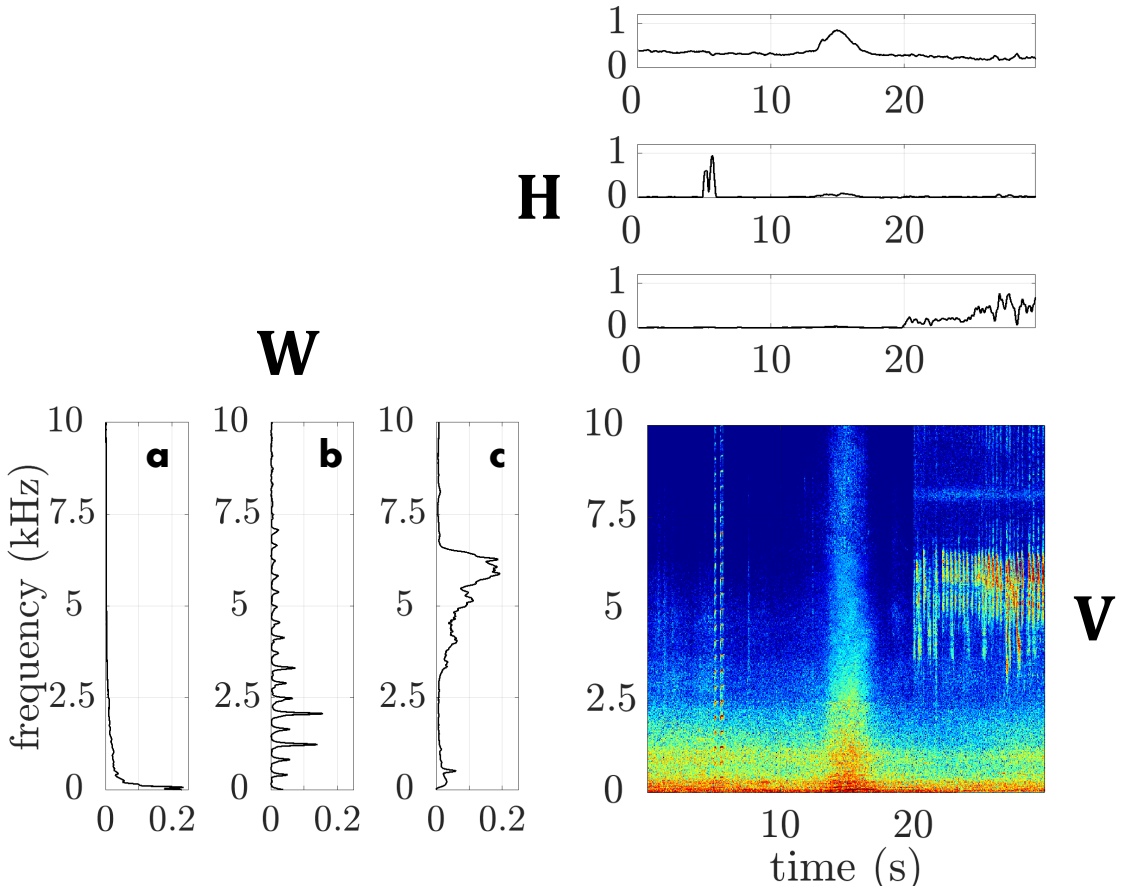


Fig. 2: NMF decomposition of ~~a~~ an urban sound mixture comprising 3 sound events (car passages, car horn and bird's whistles), **W** is composed of 3 elements too ($K = 3$) which correspond ~~in~~ to 3 audio spectra (car passages (a), car horn (b) and bird's whistles (c)).

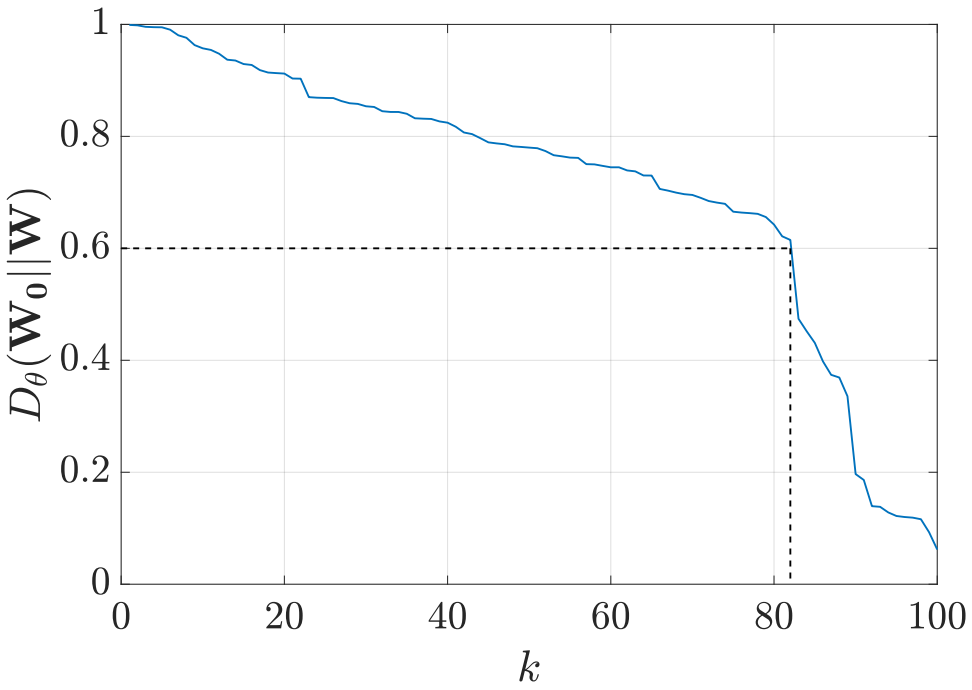


Fig. 3: $\mathbf{W}_{traffic}$ extraction from the sorted cosine similarity with a threshold $t = 0.6$. The ~~82 -nd~~ first 82 elements are considered as traffic component.

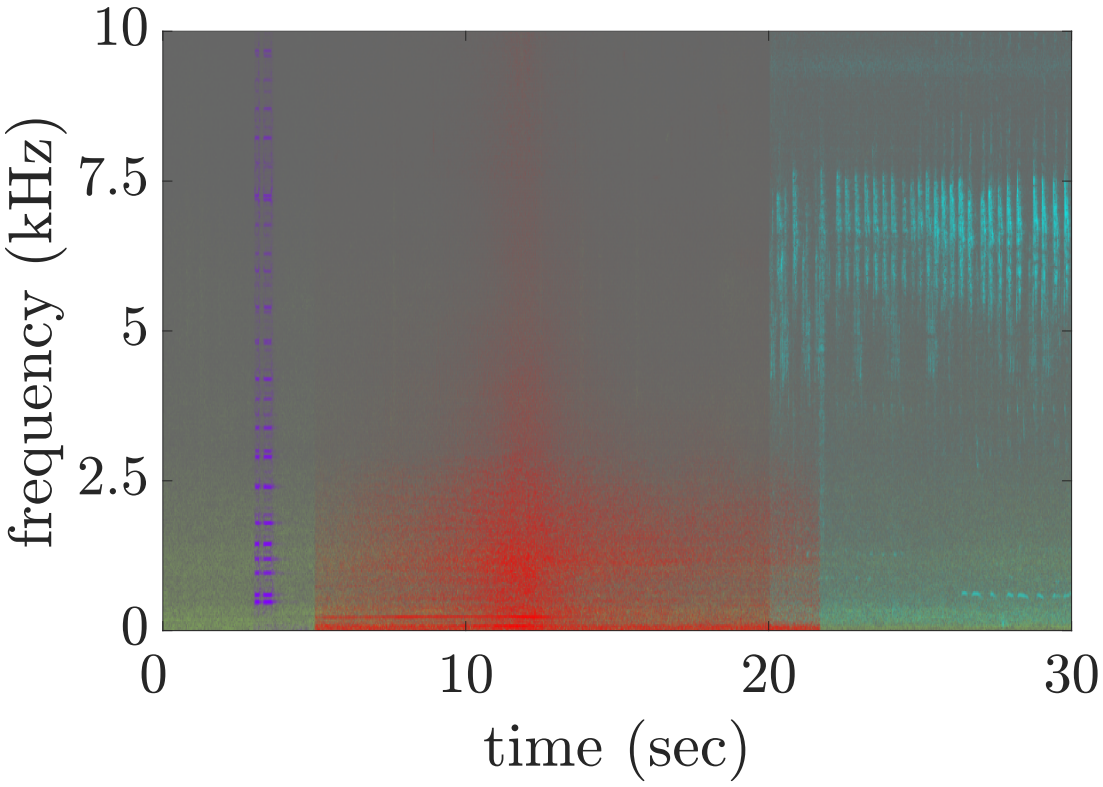


Fig. 4: Spectrogram of a sound scene created with *SimScene* software with a sound background (road traffic in green) and 3 sound events (car horn in purple, car passage in red and whistling bird in cyan).

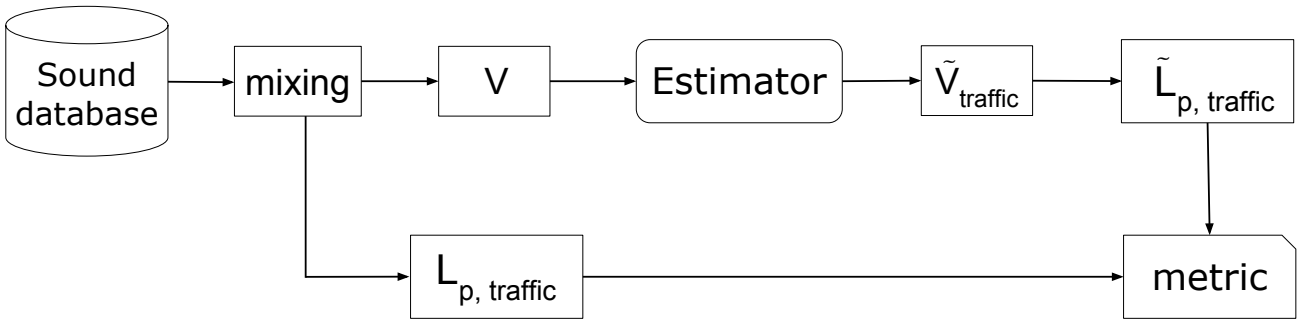


Fig. 5: Block diagram of the ~~experience~~ experiment with the urban sound scenes sound mixture step and the estimation step. The estimator may be a frequency low-pass filter or NMF.

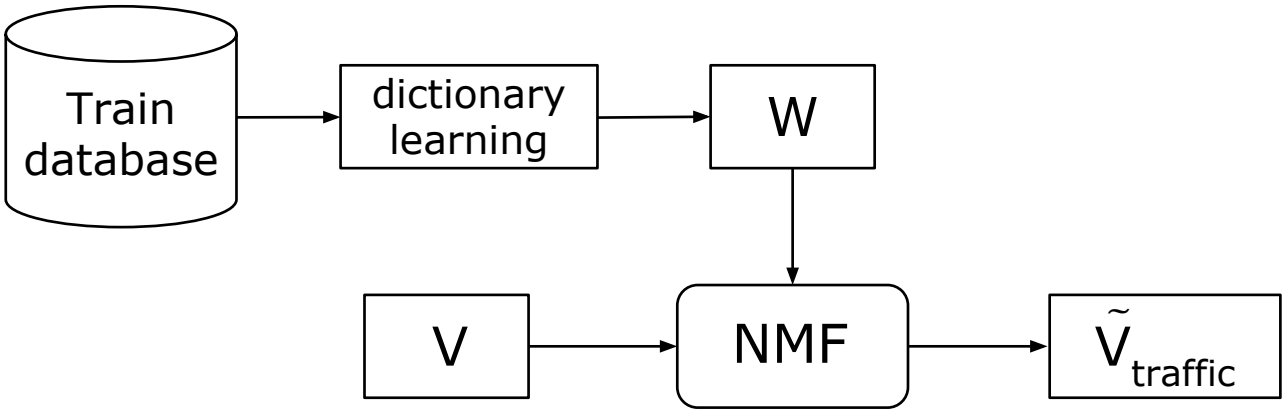


Fig. 6: Specific block diagram of the NMF estimator with the dictionary design composed from a second sound database.

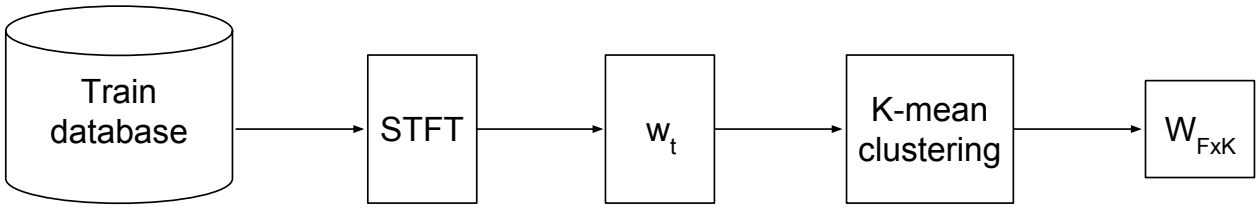
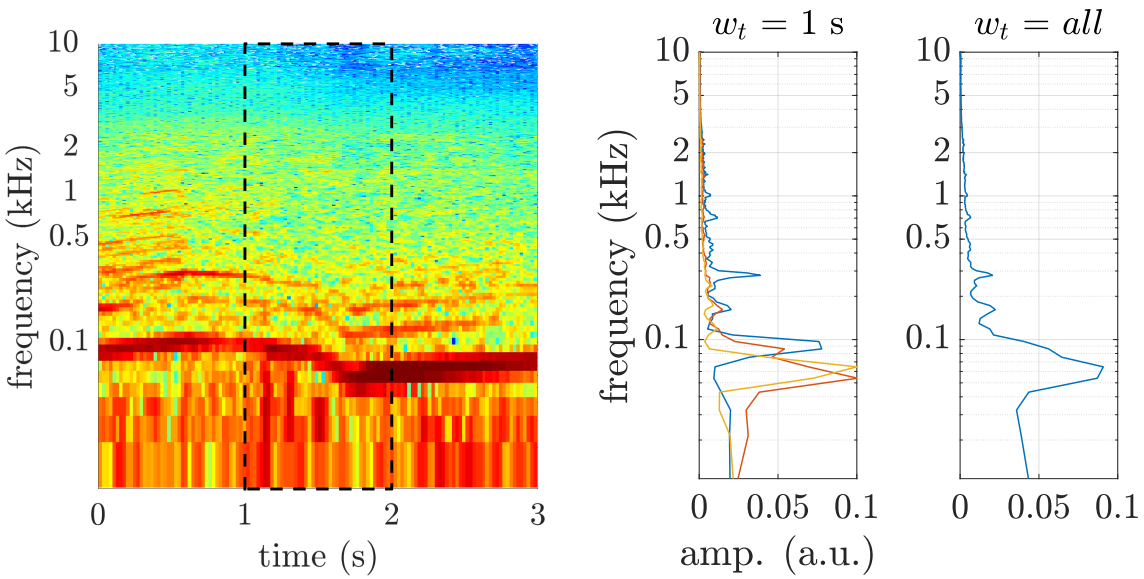
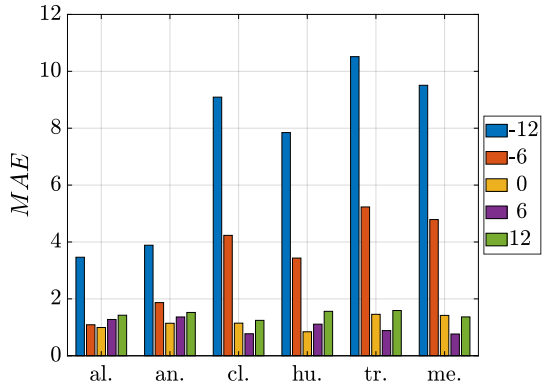


Fig. 7: Steps involved in the dictionary learning.

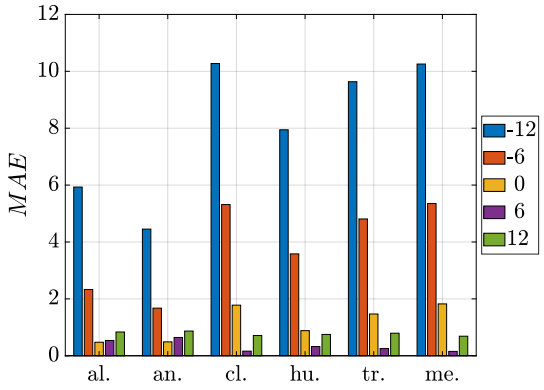


(a) (b)

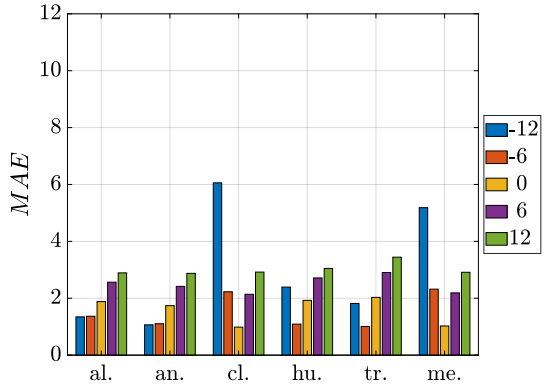
Fig. 8: Dictionary building on a 3 second extract of a car passage. In dashed lines, a 1 second w_t window (8a). With w_t = 1 second, 3 spectra are then generated and included in **W**, while for $w_t = all$, the audio file is reduced into 1 spectral vector (8b).



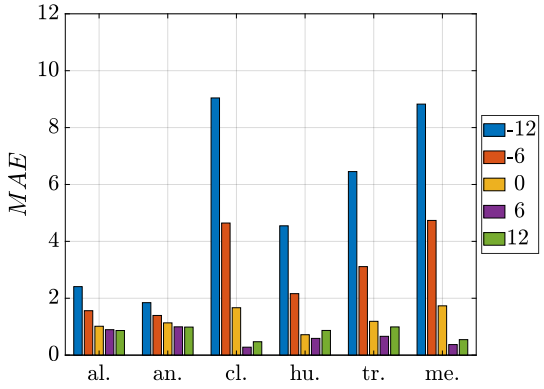
(a) MAE error for each TIR and sub-class with the frequency low-pass filter with $f_c = 500$ Hz.



(b) MAE error for each TIR and sub-class with Sup-NMF and $\beta = 2$.



(c) MAE error for each TIR and sub-class with Sem-NMF and $\beta = 1$.



(d) MAE error for each TIR and sub-class with TI-NMF, $\beta = 1$ and $t = 0.41$.

Fig. 9: MAE (dB) error for each sub-class and each TIR according to the best results with the filter (9a) and each method (Sup-NMF (9b), Sem-NMF (9c) and TI-NMF (9d)).

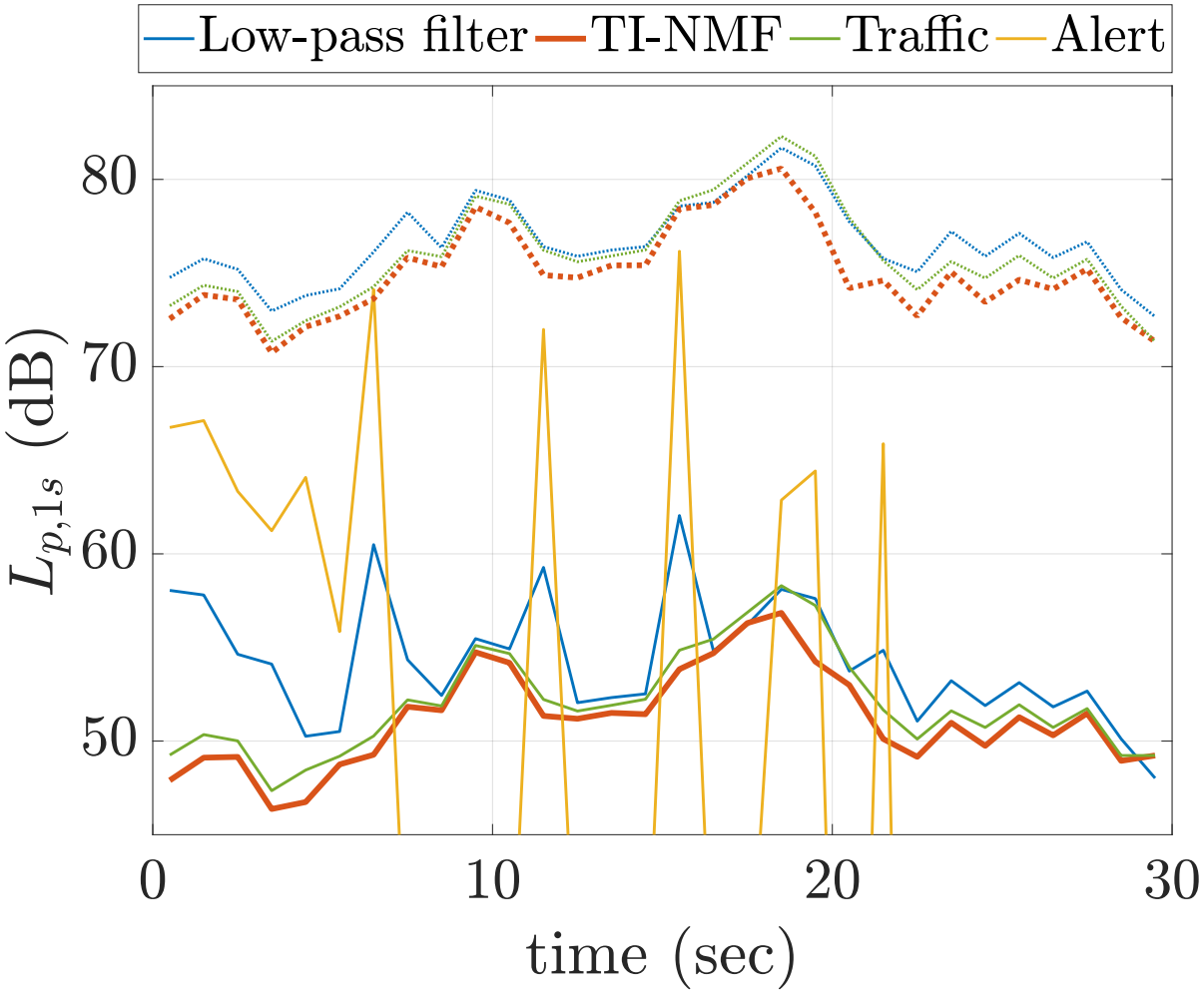


Fig. 10: Evolution of 1 second equivalent sound pressure level of an *alert* sub-class scene for 500 Hz low-pass filter and TI-NMF ($\beta = 1$, $K = 200$, $w_t = 0.5$ s, $t = 0.42$) at $TIR = -12$ dB (in full line) and $TIR = 12$ dB (in dashed line). The sound level of the *alert* sound class is null outside the y axis range.

Table 1: Summary of the different experimental factors and their modalities taken into account in the ~~frequency low-pass~~ LP filter ~~estimator~~and the NMF estimators.

experimental factors	**modalities**						**number of modalities**
sub-classes	alert	animals	climate	humans	transportation	mechanics	6
TIR (dB)	-12	-6	0		6	12	5
~~$\mathbf{f_c}$ (kHz) 0.5 1 2 5~~							
method	~~10~~ LP filter		~~20~~ Sup-NMF	~~6 Summary of the different experimental factors and their modalities taken into account in NMF estimator. **experimental factors**~~ Sem-NMF	~~**modalities**~~	TI-NMF	~~**number of modalities**~~4
~~**sub-classes** alert animals climate humans transportatic mechanics 6 **TIR** (dB~~$\mathbf{f_c}$ (kHz)	~~-12 -6 0~~ 0.5	~~6~~ 1	~~12~~ 2	5 ~~**method**~~	~~Sup-NMF Sem-NMF~~ 10	~~TI-NMF~~ 20	~~3~~ 6
$\mathbf{w_t}$ (s)	0.5		1		*all*		3
K	25		50	100		200	4
β	1				2		2
threshold **t**	from 0.30 to 0.70 with 0.01 step						41

Table 2: Best results for all the scenes according to the experimental factors β and *method* (in bold letter, the lowest error).

method	f_c (kHz)	β	**K**
filter	20	-	-
filter	0.5	-	-
Sup-NMF	-	1	~~50~~ 25
Sup-NMF	-	2	~~50~~ 25
Sem-NMF	-	1	~~100~~ 200
Sem-NMF	-	2	~~100~~ 200
TI-NMF	-	**1**	**200**
TI-NMF	-	2	~~25 all 0.54 2.20 (± 2.26)MAE error averaged on all sub-classes on each TIR for the best scenario according to each me~~

5.0.1 Experimental factors of NMF

~~Sup-NMF and Sem-NMF updates are computed for 400 iterations , which is sufficient to reach convergence. TI-NMF is performed on a lower number of iterations (60) to prevent **W** to not deviate too much from the initial dictionary~~100 iterations are performed for all NMF. The spectrogram $\mathbf{V}$ ~~, as~~ and the dictionary $\mathbf{W}$ ~~, is~~ are expressed with third octave bands ($F = 29$). This coarser method allows ~~the reduction of the matrix size and decreases~~ us to reduce the dimensionality and then decrease the computation time. ~~But, must of all~~Furthermore, by expressing the frequency axis on a log frequency axis, the low frequencies, where the traffic energy is focused, are described more finely than the high frequencies. Experimental validation consistently showed that considering third octave bands do not impact the performance of the estimator studied in this paper. But, most of all, it is a suited representation to this sound environment as this kind of representation is widely used in the urban acoustic field, compare to MFCC for instance. For TI-NMF, a preliminary study reveals that the threshold t ~~is set~~ value range can be reduce between 0.30 and 0.70~~with a~~. An increment step of 0.01 ~~. Tables ?? and~~ has been considered as being sufficiently precise. Table 1 summarize the experimental factors and their related modalities.

Considering the experimental settings derived from the different modalities of each experimental factor ~~described in Table ??~~ between the 5 levels of TIR, the 6 sub-classes and the 6 cut-off frequencies f_c, 180 settings are performed ($6 \times 5 \times 6$). For Sup-NMF and Sem-NMF, according to Table 1, 1200 associations of factors are made where the 4 levels of K are associated with $w_t \in \{0.5, 1\}$ second whereas only 2 levels of K (25 and 50) are associated with $w_t = all$, see part 4.2.1 ($6 \times 5 \times 2 \times (2 \times 4 + 1 \times 2) \times 2$). For TI-NMF, beacause of the high number of threshold t tested, 24600 combinations are computed ($6 \times 5 \times (2 \times 4 + 1 \times 2) \times 41$). In all, 25980 settings are performed.

For each setting, the estimator (frequency low-pass filter or NMF) is performed on the M scenes of a sub-class. For one sound scene, the average traffic sound level, $\tilde{L}_{p,traffic}$, of the entire scene is calculated,

$$\tilde{L}_{p,traffic} = 20 \times \log_{10}\left(\frac{p_{rms}}{p_0}\right) \tag{14}$$

where p_{rms} is the effective pressure deducted from the estimated traffic spectrogram $\tilde{\mathbf{V}}_{traffic}$ and p_0 is the reference sound pressure, $p_0 = 2 \times 10^{-5} Pa$. The A -weighting of the sound levels is not considered here as it decreases the low frequencies levels where the road traffic components are mainly present. For each setting of experimental factors, M values of $\tilde{L}_{p,traffic}$, corresponding to the M scenes, are then obtained and are compared to the M exact sound level, $L_{p,traffic}$.

5.0.2 Metrics

The performance of the road traffic sound level estimator is assessed through the calculation of one reference metric, the Mean Absolute Error (MAE) [**?**]. It expresses the quality of the long-term reconstruction of the signal and consists in the average over the M sound scenes of the absolute difference between the exact and estimated traffic sound level in dB,

$$MAE = \frac{\sum_{m=1}^{M} |L^m_{p,traffic} - \tilde{L}^m_{p,traffic}|}{M}. \tag{15}$$

~~1 second equivalent sound pressure level of an *alert* sub-class scene for 500 Hz low-pass filter and TI-NMF ($\beta = 1$, $K = 200$, $w_t = 0.5$ s, $t = 0.42$) at $TIR = -12$ dB (in full line) and $TIR = 12$ dB (in dashed line).~~

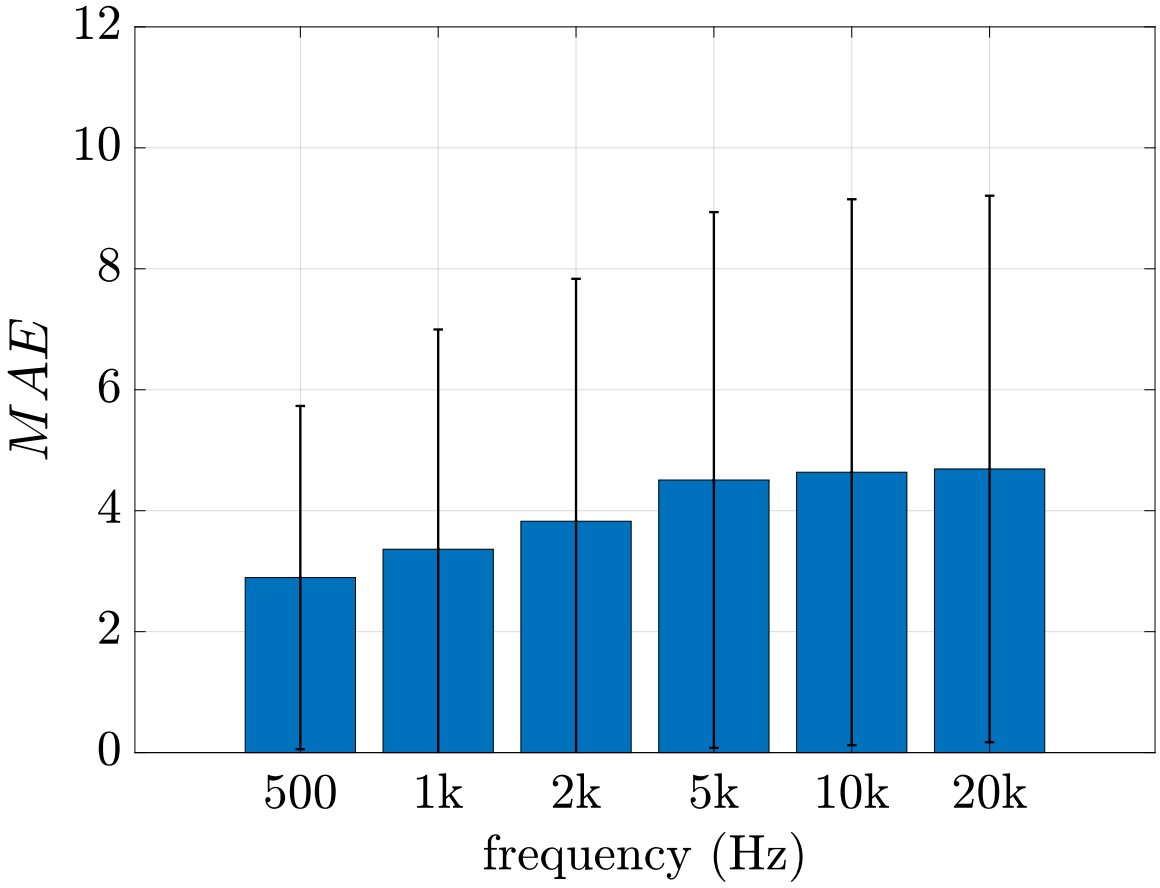


Fig. 11: Average MAE errors for the LP filter on all the corpus.

~~MAE (dB) error for each sub-class and each TIR according to the the best results with the filter (9a) and each method (Sup-NMF (9b), Sem-NMF (9c) and TI-NMF (9d)).~~ The MAE is a performance index without dimension. The use of such a metric is justified by the aim of this study that is to find the best association of experimental factors that obtain the lower mean error on all the corpus. The MAE error with logarithmic values gives an equal weight to each difference in sound level and makes it possible to be less sensitive to high energies. It is also a common practice in the environmental acoustics field when dealing with errors [?] [?].

6 Results

Table 2 summarizes, according to the 2 main factors (*method*, β), the lowest MAE error averaged on all sub-classes and all TIR (750 sound mixtures in all). For the low-pass frequency filters and each NMF approaches, the best parameter combinations are detailed according to the TIR in Table 4, and are expanded to the sub-classes in Figures 9a, 9b, 9c and 9d.

First, the errors ~~produced by the filter are detailed.~~ calculated from the traffic sound estimation of the low-pass filter on all the corpus are detailed in Figure 11. The LP filter with the cut-off frequency $f_c = 20$ kHz generates the highest error as it is equivalent to consider all the sound mixtures without distinction between traffic and others sound sources. Consequently, on Table 4 in low TIR (-12 dB and -6 dB), where traffic component is scarce, the error is more important than in high TIR (6 dB and 12 dB) where the traffic component is predominant. ~~$f_c = 500$ Hz is~~ Even if the performances for the positive TIR values are high, it has to be reminded that in practice, the TIR value is not know. In consequence,

	SUP/SEM NMF	SUP/TI NMF	SEM/TI NMF
t -value	7.81	2.84	4.60
Dof	1478	1470	1498
p -value	**5.33** $\times 10^{-15}$	**1.32** $\times 10^{-3}$	**4.57** $\times 10^{-4}$

Table 3: p -values deducted from the Student's test. In bold letters, the p -values that reject the H_0 hypothesis (p -value $\leq \alpha$).

Table 4: MAE error averaged on all sub-classes on each TIR for the best scenario according to each method.

method	filter	filter	Sup-NMF	Sem-NMF	TI-NMF
f_c (kHz)	20	0.5	-	-	-
β	-	-	2	2	1
-12	12.25 ($\pm$ 0.05)	7.39 ($\pm$ 3.00)	8.08 ($\pm$ 2.44)	2.98 ($\pm$ 2.11)	5.22 ($\pm$ 2.62)
-6	6.96 ($\pm$ 0.05)	3.44 ($\pm$ 1.65)	3.84 ($\pm$ 1.58)	1.52 ($\pm$ 0.60)	2.72 ($\pm$ 1.24)
0	3.00 ($\pm$ 0.03)	1.17 ($\pm$ 0.24)	1.15 ($\pm$ 0.62)	1.60 ($\pm$ 0.47)	1.26 ($\pm$ 0.35)
6	0.97 ($\pm$ 0.01)	1.03 ($\pm$ 0.26)	0.35 ($\pm$ 0.20)	2.49 ($\pm$ 0.30)	0.75 ($\pm$ 0.34)
12	0.26 ($\pm$ 0.00)	1.45 ($\pm$ 0.13)	0.77 ($\pm$ 0.07)	3.02 ($\pm$ 0.22)	0.83 ($\pm$ 0.23)

without a prior knowledge, this approach cannot be applied. The traffic sound level estimation error decrease with the cut-off frequency~~with~~. Finally, $f_c = 500$ Hz is the one which reach the lowest mean error obtained ($MAE =$ 2.89 ($\pm$ 2.84)~~dB~~). It is then considered as the baseline to compare the performances of NMF.

In low TIR, for *alert* and *animals*, which are sub-classes composed of higher frequencies, this filter is efficient as it removes these frequency components. For the other sub-classes where low frequency contents are present (storm for *climate*, voices in *humans*, planes, tramway and train in *transport* and ventilation noise in *mechanics*), the filter considers all the energy located in the pass-band and then does not dissociate the traffic element from the other sound sources. The errors are then nearly all superior to 4 ~~dB~~ and are overestimated. In opposite, in high TIR, the error is due to the energy removed from the traffic which has the consequence to underestimate the sound levels. The 500 Hz filter finds a balance between what is put aside in low TIR and what it is remained in high TIR.

Compared with the filter errors, the choice of some NMF approaches makes it possible to decrease the error of the road traffic sound level estimation. The supervised approach is the only method that has an average error superior to the 500 Hz filter baseline. Sem-NMF and TI-NMF have better results. The lowest average error is obtained for TI-NMF, ~~2.19~~ 2.15 ($\pm$ ~~2.18)dB~~2.10), for $\beta = 1$ and threshold $t =$ ~~0.42~~ 0.41 with the dictionary factors $K = 200$ and $w_t = 500$ ms. On the other hand, the semi-supervised approach has a higher error but a lower standard deviation ($MAE =$ 2.32 ($\pm$ ~~1.26)dB).~~ 1.15)). TI-NMF seems to be the most efficient approach on all corpus without having prior knowledge on the interfering sound class or on the TIR value.

~~According to Table~~ 4 ~~and Figures~~ 9, ~~A~~s the mean error between the best SUP-NMF ($\beta = 2$, $K = 25$, $w_t = 2$ s), SEM-NMF ($\beta = 1$, $K = 200$, $w_t = 2$ s) and TI-NMF ($\beta = 1$, $K = 200$, $w_t = 0.5$ s, $t = 0.41$) approaches are close, a Student's t-test is performed to evaluate the statistical differences between them. It is performed on the 750 estimated traffic sound levels for each couple of method (SUP/SEM NMF, SUP/TI NMF, SEM/TI NMF). The test estimates a p -value which is confronted to an significant threshold $\alpha = 5$ % that proves an H_0 hypothesis about the similarity between the distribution of the sound levels (p -value $> \alpha$). The p -values are summarized in the Table 3 with the t -values and the degrees of freedom (Dof).

For each couple of method, the ~~behavior between~~ H_0 hypothesis is rejected meaning that the distributions of the sound level estimations according to the ~~3 versions of NMF differs. In the case of Sup-NMF, it fails to improve the filtering performances despite good results in high TIR . The errors are too important for low~~ approaches are significantly different despite similar mean errors. To better understand these differences, the errors made according to the TIR ~~and this for all the~~ and for each sound environment are displayed in Table 4 and in Figure 9.