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

Jean-Rémy GLOAGUEN
Arnaud Can
LAE
Ifsttar
jean-remy.gloaguen@ifsttar.fr

Mathieu Lagrange
Jean-François Petiot
LS2N, CNRS
École Centrale de Nantes

Abstract

The advent of low cost acoustic monitoring devices raises new interesting approaches for improving the monitoring the acoustic quality of urban areas. Assuming that transportation is the major source of annoyance, state of the art approaches consider as input an estimate of the number and the speed of vehicules 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 raises 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 the non-negative factorization framework are studied in this application scenario. The task being to the best of our knowledge never been considered in the literature, we propose an exprimental protocol to validate the studied approaches that comply with standard reproducible research recommandations.

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 an estimation of the number of city dwellers exposed to high noise levels and to draw up action plans to reduce it as too long exposures to these noises can generate health problems [1]. These 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 techniques. They express L_{DEN} and L_N, which are *Day-Evening-Night* and *Night* equivalent A-weighted sound levels respectively. However, these maps introduce lot of uncertainty generated by the numerical tools [2], by the different calculation methodologies used [3][4] or even by the calculation procedure of the number of inhabitants exposed to noise [5]. In addition, the usual road traffic noise maps are static, aggregating the exposure on the two indicators L_{DEN} and L_N, thus ignoring the sound levels evolution throughout the day. Since the creation of road traffic noise maps entails long data collection and calculation times, the use of acoustic measurements could facilitate their updating or even the generation of dynamic maps [6]. These measurements can be performed at fixed stations spread all over the cities [7] [8], which would lead to the availability of the long-term evolution of the traffic noise levels. It can also be performed with mobile stations [9] [10] covering a larger area with fewer sensors but also sparse time periods.

Currently, sensor networks in cities are spread for multiple applications (air quality assessment, measurement of meteorological parameters ...), including the assessment of urban noise levels. DYNAMAP project [11] studied the deployment and feasibility of such installations. It focuses on sensor installations on specific roads at the city scale in Milan and Rome [12]. In a similar way, but reduced to few neighborhoods, the CENSE project[1] [13] aims to combine *in situ* observations, from a sensor network, and numerical data, from noise modeling, through data assimilation techniques.

If sensor networks could improve road traffic noise estimation compared with simulated maps, the issue of the correct estimation from measurements of the traffic sound level is still unsolved [7]. Indeed, the urban sound environment is a complex environment gathering lots of different sounds (car passages, voices, whistling bird, car horn ...) that can overlap. Consequently, the traffic sound level estimation based on measurements is not a trivial task. Many recent works have focused on the detection or recognition tasks of environmental sounds [14], [15], [16], [17]. A two-step process is gener-

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

ally followed : describe the audio files with a set of features (Spectrum Gravity Spectrum, 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 [18] and their applications can be found in [19], [20], [21].

The main issue in the detection or recognition tasks is the overlap of environmental sounds. Although near major roads or ring roads, traffic is predominant, there are many places where it overlaps with other sound sources that contribute significantly to the overall sound levels. Succeed to recognize two simultaneous sound sources can be very complicated. Socoro et al. propose to bypasses this issue with an Anomalous Noise Events Detector [22] which consist in detecting the sound sources from labeled recordings that are not related to the traffic component in order not to take them into account on the estimation of the traffic sound level. If the detection of the road traffic noise is good, the detection of these anomalous noise events stay a complicated task [23].

Instead of using a detection method to estimate the traffic presence, we propose in this paper to follow the blind source separation paradigm. That is, separating the contribution of the traffic from the other sources within a polyphonic scene.

One of the first to do so is the Independent Component Analysis [24]. The principle is to decompose N recorded signals to a sum of P independent sound sources weighted by linear relations. This method is most of all suited for the 'cocktail party' issue where one tries to capture a signal among noise. In an urban environment context, this method is used with acoustic microphone arrays and beamforming [25]. These approaches consider a set of microphones and allows the localization [26] or the detection [27] of sound sources from the phase shift and the distance between the microphones. However, one need a lot of microphones. Spread multiples microphones arrays in cities is then expensive (even with low cost microphones) and is time-consuming for calibration and maintenance.

A more convenient method for monophonic signal is Non-negative Matrix Factorization (NMF) [28] which consists in approximating the magnitude spectrogram of an audio file from the product of two matrices. It has been widely used in the audio domain, [29] [30] [31], and has already been employed for the source separation task of monaural signals of speech and music [32] [30]. By design, this method deals reasonably well with the overlaping sound sources as soon as the overlap can be resolved on the time/frequency plane. For the environmental sounds, the method has been used for the geolocalisation and classification of the sound environment, like in [33] where NMF is used to classify the audio files according to the 10 cities where they have been recorded.

For the sound source separation, it has also been used by Innami and Kasai in the unsupervised case [34]. After having performed NMF on simulated audio files, they realized a source separation in two steps by separating the sound background from the events first and then by separating the events between them using MFCC and time-variant gain. A k-mean clustering allows the separation of the k sound sources. This approach, compare to a simple MFCC clustering, improves the sound source separation but the authors admit their approach need the assumption that the number of sound sources is known which is unrealistic in practical case. Furthermore, in order to estimate the traffic sound level $\tilde{L}_{p,traffic}$, their approach might be difficult to be adapted it as traffic component can be both sound background and event.

Consequently, we propose in this paper to use different frameworks of Non-Negative Matrix Factorization (NMF) where traffic component is considered in its entirely whether it is a sound background or event. To validate this approach, we also consider a corpus of simulated sound scenes created from a built-up sound database composed of a high number of diverse sound samples. The use of simulated sound scenes is necessary as it offers a full control on the design of the scenes and the knowledge of the exact contribution of the traffic component ($L_{p,traffic}$) which would hardly be extracted from a recording of an urban scene.

The remaining of the paper is organized as follows. Section 2 details the technical aspects of NMF. Section 3 described on the design of the environmental sound scene corpus and the experimental protocol setup. Then Section 4 shows and discusses the results obtained during the parametric study.

2 Non-negative Matrix Factorization

2.1 Description of NMF

Non-negative Matrix Factorization is a linear approximation method introduced by Lee and Seung, [28], which can be used to approximate the spectrogram (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 spectrum and $\mathbf{H} \in \mathbb{R}^{+}_{K\times N}$ is the *activation* matrix which summarizes the temporal evolution of each element of $\mathbf{W}$. An illustrative example can be found in Figure 2.

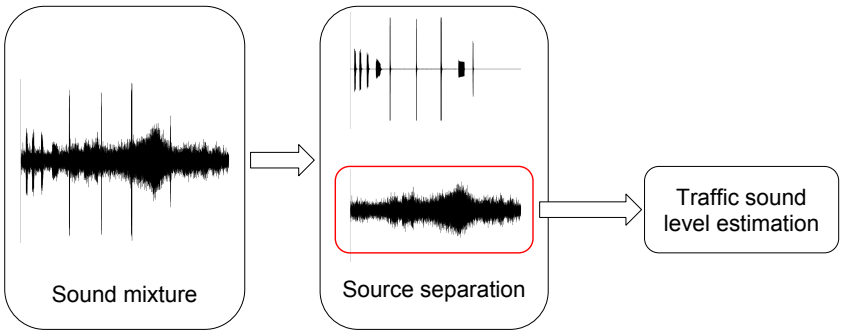


Figure 1: Bloc diagram of the source separation method

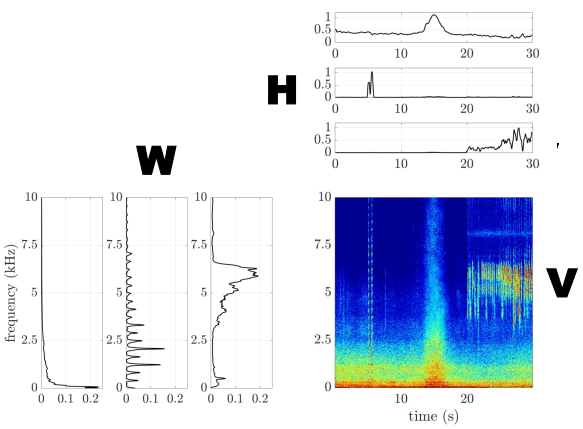


Figure 2: Example of a simple NMF for urban sound mixture, $\mathbf{W}$ and $\mathbf{V}$ are composed of 3 elements (car passages, car horn and bird's whitles)

The choice of the dimensions is often made as 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:

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

and usually belongs to the $\beta-$divergence class [35] 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 account prior knowledge on the sources, constraints (like the smoothness or the sparsness criteria [36]) can be added to the objective function.

Algorithms have been proposed to solve the minimization problem (2) iteratively such as the multiplicative update [37], the alternating least square method [38], the projected gradient [39] ... Here, the multiplicative update is chosen as it ensures non-negative results of which convergence has been proved [40].

2.2 Supervised NMF

First, supervised NMF (SUP NMF) is used: the *dictionary* includes audio spectrum of urban sound sources. In the urban environments, a lot of different sound sources present are known whose spectrum can be obtained and be a basis of $\mathbf{W}$. The *activator* are then the unknown to estimate. In the first iteration, $\mathbf{H}$ is initialized randomly, then it is updated by the generic algorithm

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

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 symbolized the Hadamard product and ratio. As in the supervised approach, the position in $\mathbf{W}$ of traffic component is known, the separation of this sound source is made by extracting the related basis and activators,

$$\tilde{\mathbf{V}}_{traffic} = \left[\mathbf{WH}\right]_{traffic}. \tag{6}$$

2.3 Semi-supervised NMF

One of the main issue with the supervised approach is the generalization issue: how to be adapted to different sound mixtures with a fixed dictionary ? To better

take into account the diverse nature of urban scenes, semi-supervised NMF (S-S NMF)[41] can be useful as it has been proposed to offer more flexibility. This method consists in composing the *dictionary* with a fixed part $\mathbf{W_s} \in \mathbb{R}^{+}_{F\times K}$, composed in our case of road traffic spectrum, and with a mobile part, $\mathbf{W_r} \in \mathbb{R}^{+}_{F\times J}$ with $J << K$, that is updated. Here, $J = 2$. The aim is to include in $\mathbf{W_r}$ the elements that are not related with the traffic. The problem (1) becomes

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

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 S-S NMF for speech denoising from background noise or musical content can be found in [42] and [43].

2.4 Thresholded initialized NMF

A last approach is tested based on unsupervised NMF. Usually, $\mathbf{W}$ is learnt with the help of a learning corpus by initiated it randomly. Here, as the concerned sound source is known and audio samples of car passages are available, a initial dictionary, $\mathbf{W_0}$, is learnt by converting the audio files in the spectra domain; see part 3.2.1. 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 with *a priori* knowledge.

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

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

$$D_\theta\left(\mathbf{W_0}||\mathbf{W}\right) = \frac{\mathbf{W}.\mathbf{W_0}}{||\mathbf{W}||.||\mathbf{W_0}||}. \tag{10}$$

$D_\theta\left(\mathbf{W_0}||\mathbf{W}\right) = 1$ means that the elements are identical (the k-th element of $\mathbf{W}$ is then considered as traffic element) whereas $D_\theta\left(\mathbf{W_0}||\mathbf{W}\right) = 0$ means that the elements are significantly different. This measure is bounded between 1 and 0 and is an invariant scale estimation of the similarity. Then, $D_\theta\left(\mathbf{W_0}||\mathbf{W}\right)$ is sorted in descending order. The elements in $\mathbf{W}$ that can belong to $\mathbf{W}_{traffic}$ are then selected by a *hard thresholding* method. It is defined as:

$$\mathbf{W}_k \in \mathbf{W}_{k,traffic} \quad \text{iff} \quad D\left(\mathbf{W}_{0,k}||\mathbf{W}_k\right) > t \tag{11}$$

An illustrative example can be see in Figure 3.

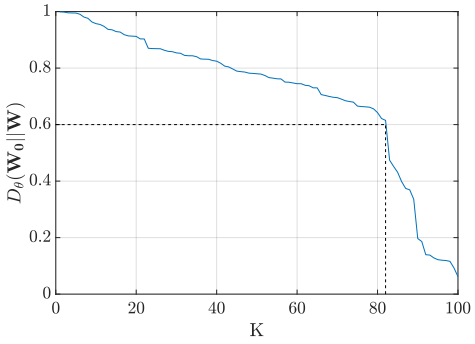


Figure 3: Example of the $\mathbf{W}_{traffic}$ extraction from the sorted cosine similarity with a threshold $t = 0.6$. The 82-nd first elements are considered as traffic component.

This approach is named *Thresholded inititialized NMF* (TI NMF). Other thresholding methods as the *soft* [44] and the *firm* [45] and multiples way to display the similarity through a sigmoïd or a Radial Basis Function have been investigated. A fast parametric study has revealed that the *hard* thresholding method with a linear representation of the similarity according to K, as in Figure 3, was the best way to get better performances.

3 Experimental protocol

In order to validate the usefulness of considering NMF framework to estimate the road traffic noise level, one need to have a reference level. It can hardly be measured or even annotated from real life recordings. Thus, simulated sound scenes are used to assess the performance of the proposed NMF. This offers a controlled framework to design specific sound environments in which all the traffic component is known. Then, the estimated road traffic sound levels with the method can be compared to the real ones, introduced within each simulated sound scene.

3.1 Environmental sound scene corpus

A corpus is designed with the *SimScene* software[2]. *SimScene* [46] is a simulator that creates sound scenes in a .wav format by summing audio samples that come from an isolated sound database.

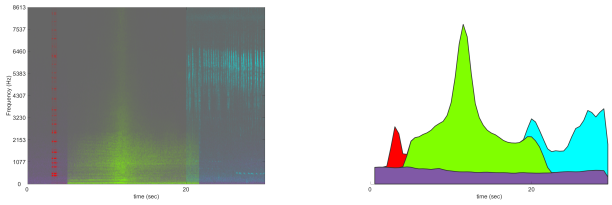
This database is divided in two categories: *i*) the *event* category which are the brief sounds (from 1 to 20 seconds) that are considered as salient including 245 sound event samples divided in 19 sound classes (*ringing bell, birds, sweeping broom, car horn, car passages, hammer, drill, coughing, barking dog, rolling suitcase, closing door, plane, siren, footstep, storm, street noise, train, tramway, truck and voice*) and *ii*) the *background* 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 belong to this category divided in 9 sound classes (*birds, construction site noise, crowd, park, rain, children playing in schoolyard, constant traffic noise, ventilation, wind*). The sound class *car passages* comes from 60 recordings of 2 cars (Renault Megane and Renault Senic) made on the Ifsttar's runway on different speeds with multiple gear ratio. The other audio files have been found online (*freesound.org*) and within the *UrbanSound8k* database [47]. Each sound class is composed of multiples samples (*bird01.wav*, *bird02.wav* ...). The software allows the user to control some 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 to know its exact contribution as well as a text file that summarizes the time presence of all the events.

This database enables creating realistic urban sound scenes from the road traffic point of view [48]. A sound mixing corpus is composed of 6 sub-corpus 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 birds), *climate* (wind, rain), *humans* (crowd noise and voice), *mechanics* (different metallic and construction site noises) and *transportation* (train, tramway and plane). In each file, traffic component is present as the sum of the background and event traffic sounds and is mixed with the other sound classes. The sound classes that are not related to the traffic component are summed up as the *interfering* sound class. 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).

$$TIR = L_{p,traffic} - L_{p,interfering} \tag{12}$$

with the *Traffic Interference Ratio* $TIR = [-12, -6, 0, 6, 12]$. When $TIR = -12$, the traffic component is then less present than the interfering class. When $TIR = 12$ it is the opposite: the trafffic class is louder than the interfering class. The total number of scenes designed is 750 (6 sub-corpus $\times$ 25 scenes $\times$ 5 TIR values).

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



(a) (b)

Figure 4: Example of a simple scene created with *SimScene* software ((4a) the spectrogram, (4b), the time domain) with a sound background (road traffic in purple) and 3 sound events (car horn in red, car passage in green and whistling bird in blue):

3.2 Experiment

The experiment consists in estimating the traffic road sound level of the 6 environmental sound sub-corpus (*alert* (al), *animals* (an), *humans* (hu), *climate* (cl), *mechanics* (me), *transportation* (tr)) and for 5 TIR ([-12 -6 0 6 12] dB). The spectrogram $\mathbf{V}$ of each sound is built with a window size $w = 2^{12}$ with a 50 % overlap and a number of point $nfft = 2^{12}$ ($N = 644$).

The first estimator to determine the traffic sound level is a frequency low-pass filter which depends only on the cut-off frequencies f_c = [500 1k 2k 5k 10k 20k] Hz (see Figure 5a). The spectrogram $\mathbf{V}$ is filtered and the remaining energy is then considered as traffic component (eq. 13),

$$\tilde{\mathbf{V}}_{traffic} = \mathbf{V}_{f_c}. \tag{13}$$

The second estimator is the proposed scheme, based on several flavors of the NMF framework. Multiples experimental factors are involved here between the dictionary learning and NMF (see Figure 5b).

Experimental factors	value					
K	25	50	100	200		
$\mathbf{w_t}$ (s)	0.5		1		*all*	
TIR (dB)	-12	-6	0	6	12	
sub-classes	alert	animals	climate	humans	transportation	mechanics
β	1			2		
$\mathbf{f_c}$ (kHz)	0.5	1	2	5	10	20
method	filter		SUP NMF	S-S NMF	TI NMF	
t	0.20:0.01:0.70					

Table 1: Summary of the different experimental factors taken into account in the frequency low-pass filter and NMF process and their values for the estimation of the traffic sound level

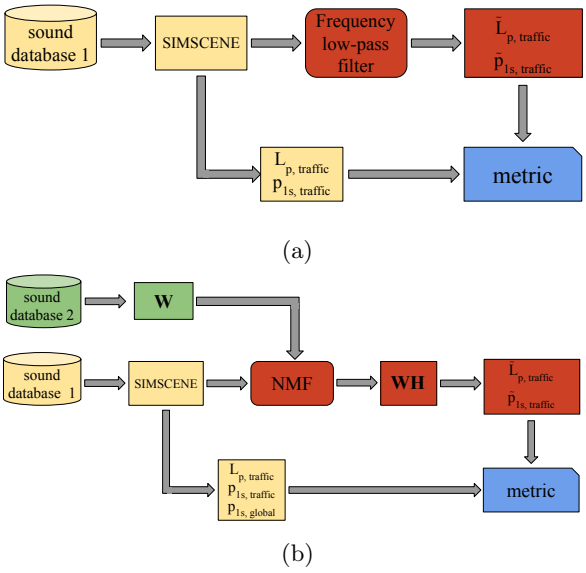


Figure 5: Block diagrams summed up the different step of the process for the frequency low-pass filter (a) and for NMF (b)

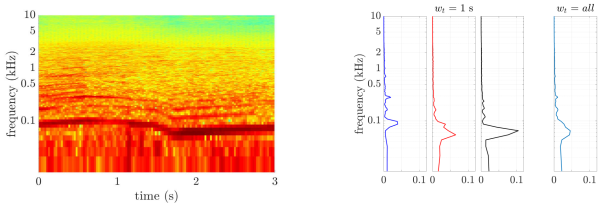


Figure 6: Example of the dictionary building with a 3 second extract of a car passage (6a). With $w_t = 1$ second, 3 spectrum are generated and included in $\mathbf{W}$ while for $w_t = all$, the audio file is resume to 1 spectra (6b).

3.2.1 Dictionary building

The dictionary is built from a second sound database dedicated specifically to this task. It is composed of 53 audio files of passing cars. These records have been made on the Ifsttar's runway too with the same experimental conditions that the records made for the *SimScene* database but with two different cars (Dacia Sandero and Renault Clio). First, for each audio file, its spectrogram is calculated with fixed parameters (w, 50 % overlap, $nfft$). Then time/frequency windows of $F \times w_t$ dimension are applied without overlapping on the spectrogram in order to consider several spectrum for each audio file. w_t is fixed at $wt = [0.5\ 1]$ second. In each window, the root mean square value is calculated on each frequency bin to reduce the windowed spectrogram in one spectra of $F \times 1$ dimension. A special case is added where the root mean square of *all* the spectrogram is applied (each audio file generates one element k of $\mathbf{W}$). An example that illustrates the process can be found on Figure 6 on a 3 second extract of the spectrogram of a car passage ; see Figure 6a. 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 to one element ; see Figure 6b.

Since the number of elements given by this processing can be high, in order to reduce the computational time

and delete redundant information, a K-means clustering is applied to reduce the number of spectrum to $K = [25, 50, 100, 200]$.

The obtained dictionary is expressed with third octave bands and 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 values.

3.2.2 Experimental factors of NMF

SUP and S-S NMF are performed for 400 iterations which is sufficiently enough to get a stabilized reconstruction. TI NMF is performed on a lower number of iteration (60) to prevent $\mathbf{W}$ to not deviate too much from the initial dictionary.

The spectrogram $\mathbf{V}$ and the dictionary $\mathbf{W}$ are expressed with third octave bands ($F = 29$). This coarser method allows the matrix reduction and decrease the computation time. But, must of all, 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. The case of a linear frequency axis has been investigated and offer similar results which were not better.

For TI NMF, the threshold is define between 0.20 and 0.70 with a 0.01 step. Table 1 summarizes the experimental factors and their related values.

3.2.3 Metrics

The performances of the two estimators (filter and NMF) of the road traffic sound level are assessed through the calculation of one 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,

$$MAE = \frac{\sum_{m=1}^{M} |L^m_{p,traffic} - \tilde{L}^m_{p,traffic}|}{M}. \qquad (14)$$

In all, according Table 1, 24540 settings are performed between the different form of the dictionary $\mathbf{W}$ and the multiple experimental factors taken into account by NMF. For the filter estimator, between the TIR, the sub-classes and f_c, 180 settings are performed whereas for SUP and S-S NMF, 3780 associations of factors are made. For TI NMF, 8400 combinations can be calculated.

4 Results

Table 2 summarized, according to the 2 main factors (*method*, β), the MAE error averaged on all sub-classes and all TIR (750 sound mixtures in all). For the low-passa frequency filters and each NMF approaches, the best parameter combinations are detailed according to the TIR in Table 3, and are expand to the sub-classes in Figures 8a, 8b, 8c and 8d.

First, the errors produced by the filter are detailed. $f_c = 20$ kHz is equivalent to consider all the sound mixtures without distinction between traffic and others sound sources. Consequently, 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 the cut-off frequency with the lower mean error obtained. It is then the first baseline to use to compare the performances of NMF.

In low TIR, for *alert* and *animals*, 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 do not dissociate the traffic element from the other sound sources. The errors are then nearly all superior to 4 dB and are overestimate ; see Figure 7a. 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 ; see Figure 7b. The 500 Hz filter finds a balance between what it is put aside in low TIR and what it is remained in high TIR.

Compare to the filter errors, the choice of some NMF approaches make it possible to decrease 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. Finally, it is the most sophisticated methods (S-S and TI NMF) that have better results. The lowest average error is obtained for TI NMF for $\beta = 1$ and threshold $t = 0.42$ with the dictionary factors $K = 200$ and $w_t = 500$ ms. On the other hand, the semi-supervised approach has a higher error but yet with a lower standard deviation.

According to Table 3 and Figures 8, the behavior between 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 error are too important for low TIR and this for all the sub-classes. This method reveals to be to rigid as $\mathbf{W}$ is composed of fixed traffic spectrum. 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. The case of a scene of the *alert* sub-class is presented in Figure 9a. Here when the car horn sounds,

method	β	**K**	$\mathbf{w_t}$ (s)	**t**	MAE **(dB)**
filter 20 kHz					4.69 (± 4.52)
filter 0.5 kHz					2.89 (± 2.84)
SUP NMF	1	50	0.5		3.44 (± 3.70)
SUP NMF	2	50	0.5		3.02 (± 3.33)
S-S NMF	1	100	0.5		2.33 (± 1.10)
S-S NMF	2	100	0.5		2.32 (± 1.26)
TI NMF	1	200	0.5	0.42	**2.19 (± 2.18)**
TI NMF	2	25	all	0.54	2.20 (± 2.26)

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

method	filter	filter	SUP NMF	S-S NMF	TI NMF
f_c (kHz)	20	0.5	20	20	20
β			2	2	1
-12	12.25 (± 0.05)	7.36 (± 3.00)	8.65 (± 1.88)	3.88 (± 1.52)	5.35 (± 2.71)
-6	6.96 (± 0.05)	3.44 (± 1.65)	4.22 (± 1.27)	1.37 (± 0.71)	2.82 (± 1.30)
0	3.00 (± 0.03)	1.17 (± 0.24)	1.34 (± 0.56)	1.11 (± 0.25)	1.26 (± 0.35)
6	0.25 (± 0.06)	1.03 (± 0.26)	0.26 (± 0.10)	2.25 (± 0.19)	0.70 (± 0.32)
12	0.26 (± 0.00)	1.45 (± 0.13)	0.64 (± 0.06)	2.96 (± 0.21)	0.84 (± 0.24)

Table 3: MAE error averaged on all sub-classes on each TIR for the best scenario according to each method

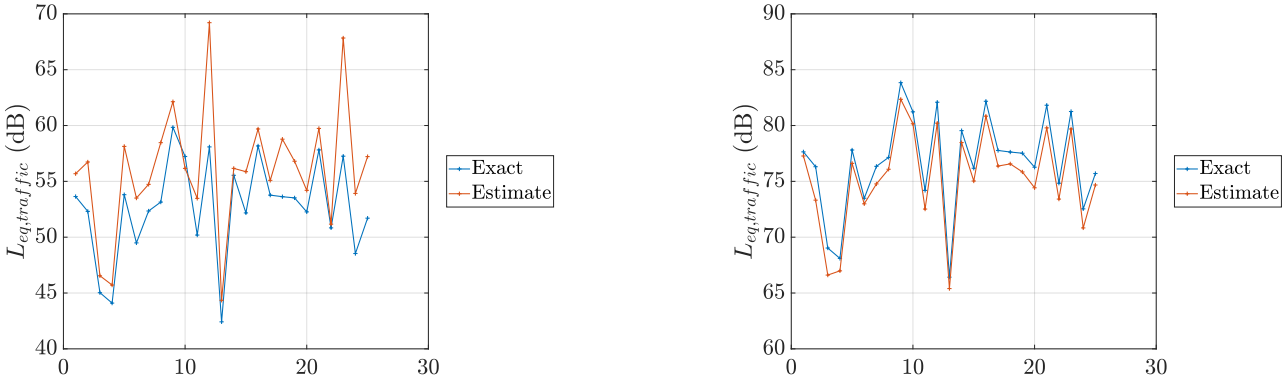
some elements of $\mathbf{W}$ are activated and are then considered as traffic component generating a wrong estimation of the sound level. This behavior disappear when the traffic component become predominant to the *interfering* class, see Figure 9b. Thus composed the dictionary of only fixed traffic spectrum is not a sufficient way to estimate correctly the traffic sound level, $\tilde{L}_{p,traffic}$.

Consequently, with the add 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 as it can be see in Figure 8c. The $\mathbf{W_r}$ composition, in the case of an *alert* sub-class, is displayed in Figure 10a and this shows that the interfering class is easily integrated. The first element is mainly composed of high frequency bands which correspond to the car horns of the scene. This composition impacted directly the traffic sound level estimation as it can be seen in Figure 9a where the traffic basics are no longer activated when the car horn sounds.

However, the degrees of freedom of S-S NMF are restrictive for high TIR as the errors exceed 2 dB for all sub-classes and increase with TIR = 6 dB and 12 dB. Indeed, in order to reduce $D(\mathbf{V}||\mathbf{WH})$, without constraint, S-S NMF is free to include traffic components in $\mathbf{W_r}$. Consequently, this behavior decreases the quality of the reconstruction of the traffic component. In Figure 10b for TIR = 12 dB, the high frequency components of car horn have disappeared for the benefit of low frequency content. Consequently, the traffic sound level estimation is then underestimate as in Figure 9b.

Finally, TI NMF with a threshold fixed at $t = 0.54$ and $\beta = 2$, offers the lowest average results (Table 2). If, according to each TIR, this method does not propose the lowest error, it succeeds to be a compromise between the two others NMF. Unlike SUP NMF, where $\mathbf{W}$ is fixed, and S-S NMF, where only $\mathbf{W_r}$ is updated, TI NMF updates $\mathbf{W}$ entirely to be adjusted to the scene and 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. In Figure 11, the similarity $D_\theta(\mathbf{W_0}||\mathbf{W})$ is displayed for 3 sub-classes for TIR = [-12, 12] dB.

For low TIR, as the traffic sound class is not predominant, the final dictionary differs a lot from the $\mathbf{W_0}$. With the thresholding, only a reduce number of basis vectors are considered as traffic components. In comparison to supervised results, this approach reduce significantly the error for the *human* and *transport* sub-classes. However, for *climate* and *mechanics*, the error stay important as these interfering classes have similar spectrum. On 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 and then decrease the error ($<$ 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. With



(a) (b)

Figure 7: Global sound levels of the traffic estimated by the frequency low-pass filter with $f_c = 500$ Hz for the sub-classes *alert*: at $TIR = -12$ (7a) and at TIR (7b).

a low threshold, it is possible to decrease the error (for $TIR = 12$ dB, the average error on all the sub-classes is 0.22 ($\pm$ 0.08) dB with $t = 0.30$).

5 Conclusion

In this work, NMF, a source separation method, was used to estimate the road traffic sound level in urban sound mixtures. It is an adapted approach to these sound environments because it easily takes into account the overlap between the multiple sound sources present in the cities and 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 sound, 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 new approach, named Threshold Initialized NMF, where $\mathbf{W}$ is initialized with road traffic spectrum, updated and where the traffic elements are extracted by hard thresholding, proposes the lowest average error. It allows a compromise between the other two approaches where a prior knowledge of the targeted sound source is given as in the SUP NMF, and offers an adaptability to the different sound mixtures by updating $\mathbf{W}$ entirely.

This method has the advantage to keep a low calculation cost and then can be implemented in embedded sensors. Furthermore, this approach stay adapted for other sound sources, as whistling bird, depending on what spectrum are put in $\mathbf{W}$ and can be adapted for other application than the road traffic sound level estimation. Finally, the add of constraints as sparsness [49] and smoothness [36] can easily be implemented and taken into account.

In parallel, this experience allowed the design of a large sound database composed of multiples samples of urban sounds and sound mixtures that are made available[3] for research communities dedicated to the detection, separation and recognition tasks of urban sound sources. The experience has been lead with the Matlab software. In order to be a reproducible experience, all the programs[4] used for this experience are available online.

References

[1] World Health Origanization. Burden of disease from environmental noise. Quantification of healthy life years lost in Europe. `http://www.euro.who.int/en/home/copyright-notice`. visité le 24/08/2017.

[2] H. Van Leeuwen and S. Van Banda. Noise mapping - State of the art - Is it just as simple as it looks? *EuroNoise*, 2015.

[3] O. Leroy, B. Gauvreau, F. Junker, E. De Rocquigny, and M. Berengier. Uncertainty assessment for outdoor sound propagation. In *20th International Congress on Acoustics, ICA 2010*, page 7p, France, August 2010.

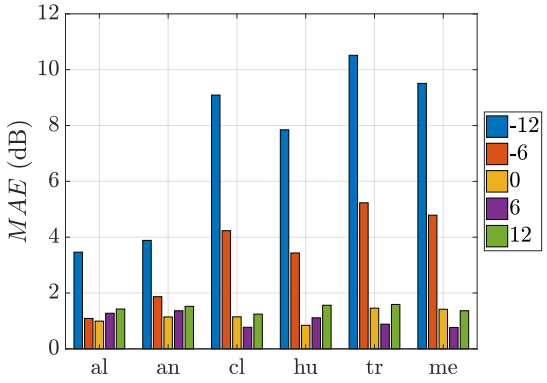
[4] N. Garg and S. Maji. A Critical Review of Principal Traffic Noise Models: Strategies and Implications. *Environmental Impact Assessment Review*, 46, April 2014.

[5] E. A. King, E. Murphy, and H. J. Rice. Implementation of the EU environmental noise directive: lessons from the first phase of strategic noise mapping and action planning in Ireland. *Journal of Environmental Management*, 92(3):756–764, March 2011.

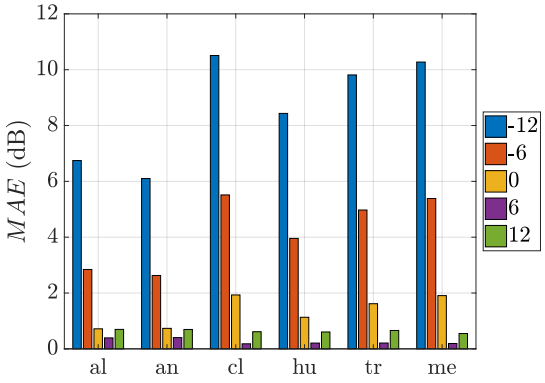
[6] W. Wei, T. Van Renterghem, B. De Coensel, and D. Botteldooren. Dynamic noise mapping: A map-based interpolation between noise measurements with high temporal resolution. *Applied Acoustics*, Complete(101):127–140, 2016.

[3] `https://sandbox.zenodo.org/record/176695#.Wk4ow3kiGos`

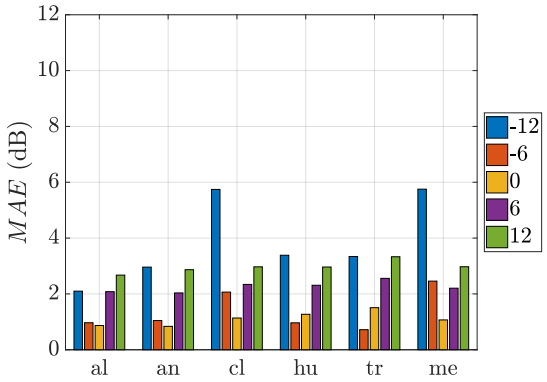
[4] `https://github.com/jean-remyGloaguen/article2017EstimationAmbiance`



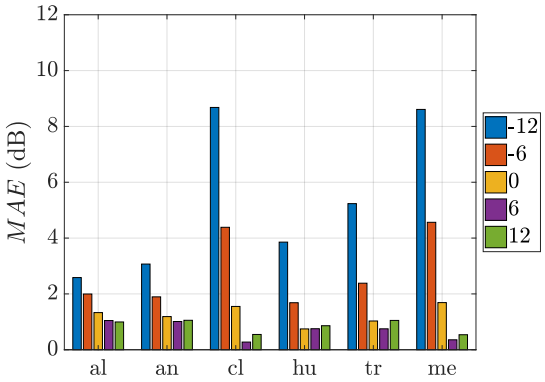
(a) Frequency low-pass filter with $f_c = 500$ Hz



(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 S-S NMF and $\beta = 2$



(d) MAE error for each TIR and sub-class with TI NMF, $\beta = 1$ and $t = 0.42$

Figure 8: MAE error for each sub-class and TIR according to the the best results with the filter (8a) and each method (SUP (8b), S-S (8c) and TI (8d NMF))

[7] P. Mioduszewski, J. A. Ejsmont, J. Grabowski, and D. Karpinski. Noise map validation by continuous noise monitoring. *Applied Acoustics*, 72(8):582–589, july 2011.

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

[9] A. Can, T. Van Renterghem, and D. Botteldooren. Exploring the use of mobile sensors for noise and black carbon measurements in an urban environment. In Société Française d'Acoustique, editor, *Acoustics 2012*, Nantes, France, April 2012.

[10] D. Manvell, L. Ballarin Marcos, H. Stapelfeldt, and R. Sanz. Sadmam-combining measurements and calculations to map noise in madrid. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2004, pages 1998–2005. Institute of Noise Control Engineering, 2004.

[11] S. Xavier, S. J. Claudi, A. Francesc, B. Patrizia, and al. DYNAMAP – Development of low cost sensors networks for real time noise mapping. *Noise Mapping*, 3(1), May 2016.

[12] P. Bellucci, L. Peruzzi, and G. Zambon. LIFE DYNAMAP project: The case study of Rome. *Applied Acoustics*, Part B(117):193–206, 2017.

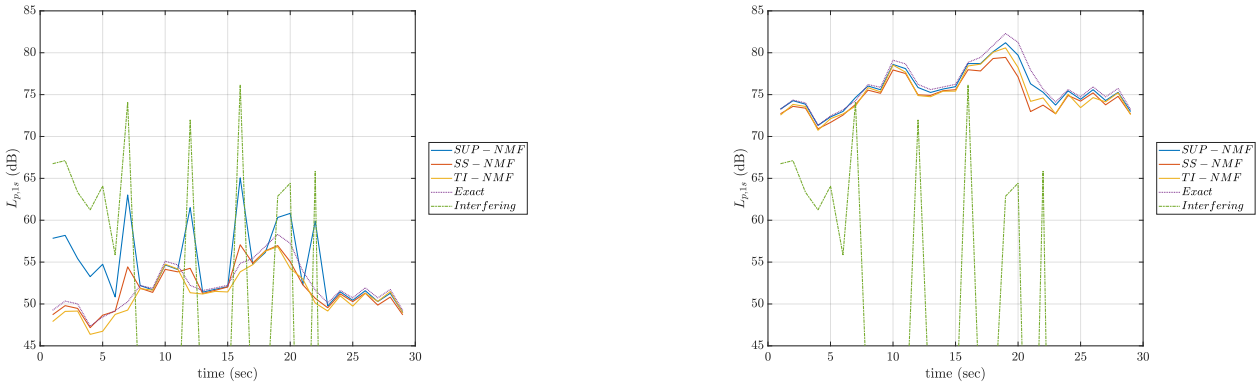
[13] J. Picaut, A. Can, J. Ardouin, P. Crépeaux, T. Dhorne, D. Écotière, M. Lagrange, C. Lavandier, V. Mallet, C. Mietlicki, et al. Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling. *The Journal of the Acoustical Society of America*, 141(5):3808–3808, 2017.

[14] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen. Sound event detection in multisource environments using source separation. In *in Workshop on Machine Listening in Multisource Environments, CHiME2011*, 2011.

[15] B. Defreville, F. Pachet, C. Rosin, and P. Roy. Automatic Recognition of Urban Sound Sources. Audio Engineering Society, 2006.

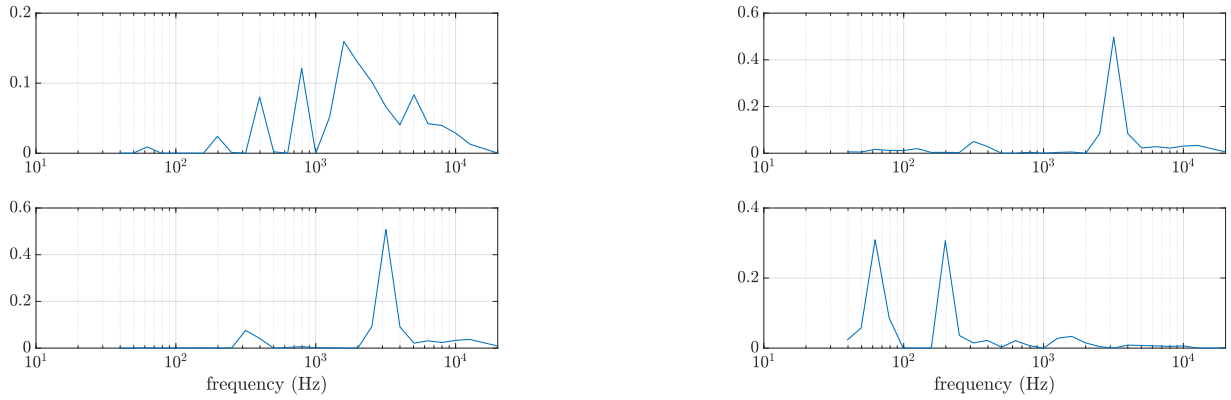
[16] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini. Automatic sound detection and recognition for noisy environment. In *2000 10th European Signal Processing Conference*, pages 1–4, September 2000.

[17] S. Chu, S. Narayanan, and C. C. J. Kuo. Environmental Sound Recognition With Time-Frequency Audio Features.



(a) (b)

Figure 9: 1 second equivalent sound pressure level of an *alert* sub-class scene for : at $TIR = -12$ (9a) and at TIR (9b).



(a) (b)

Figure 10: 2 elements of $\mathbf{W_r}$ for an *alert* scene for TIR = -12 (10a) and TIR = 12 (10b)

IEEE Transactions on Audio, Speech, and Language Processing, 17(6):1142–1158, August 2009.

[18] M. Cowling and R. Sitte. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895–2907, November 2003.

[19] G. Shen, Q. Nguyen, and J. Choi. An Environmental Sound Source Classification System Based on Mel-Frequency Cepstral Coefficients and Gaussian Mixture Models. *IFAC Proceedings Volumes*, 45(6):1802–1807, May 2012.

[20] F. Beritelli and R. Grasso. A pattern recognition system for environmental sound classification based on MFCCs and neural networks. In *2008 2nd International Conference on Signal Processing and Communication Systems*, pages 1–4, December 2008.

[21] L. Couvreur and M. Laniray. Automatic Noise Recognition in Urban Environments Based on Artificial Neural Networks and Hidden Markov Models. In *The 33 rd International Congress and Exposition on Noise Control Engineering*, Prague, August 2004.

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

[23] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[24] P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, April 1994.

[25] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano. Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, 2003:1135–1146, 2003.

[26] D. Mennitt and M. Johnson. Multiple-array passive acoustic source localization in urban environments. *The Journal of the Acoustical Society of America*, 127(5):2932–2942, 2010.

[27] R. Leiba, F. Ollivier, J. Marchal, N. Misdariis, R. Marchiano, et al. Large array of microphones for the automatic recognition of acoustic sources in urban environment. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 255, pages 2662–2670. Institute of Noise Control Engineering, 2017.

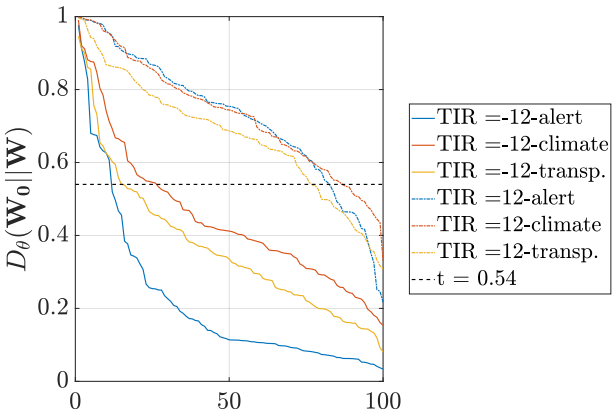


Figure 11: Example of the similarity $D_\theta(\mathbf{W_0}||\mathbf{W})$ for different sub-classes and for two TIR (-12 dB and 12 dB) and with the threshold $t = 0.54$

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

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

[30] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4029–4032, March 2008.

[31] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen. Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155, April 2015.

[32] B. Wang and M. D. Plumbley. Musical audio stream separation by non-negative matrix factorization. *Proc. DMRN summer conf*, pages 23–24, 2005.

[33] A. Kumar, B. Elizalde, and B. Raj. Audio Content based Geotagging in Multimedia. *arXiv:1606.02816 [cs]*, June 2016. arXiv: 1606.02816.

[34] I. Satoshi and K. Hiroyuki. NMF-based environmental sound source separation using time-variant gain features. *Computers & Mathematics with Applications*, 64(5):1333–1342, 2012.

[35] C. Févotte, N. Bertin, and J. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[36] T. Virtanen. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, March 2007.

[37] D. Lee and H. Seung. Algorithms for Non-negative Matrix Factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.

[38] A. Cichocki and R. Zdunek. Regularized Alternating Least Squares Algorithms for Non-negative Matrix/Tensor Factorization. In *Advances in Neural Networks – ISNN 2007*, Lecture Notes in Computer Science, pages 793–802. Springer, Berlin, Heidelberg, June 2007.

[39] C. J. Lin. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*, 19(10):2756–2779, October 2007.

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

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

[42] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller. Real-time speech separation by semi-supervised nonnegative matrix factorization. *Latent Variable Analysis and Signal Separation*, pages 322–329, 2012.

[43] F. Weninger, J. Feliu, and B. Schuller. Supervised and semi-supervised suppression of background music in monaural speech recordings. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 61–64. IEEE, 2012.

[44] D. L. Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995.

[45] M. Fornasier and H. Rauhut. Iterative thresholding algorithms. *Applied and Computational Harmonic Analysis*, 25(2):187–208, 2008.

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

[47] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22st ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014.

[48] J.-R. Gloaguen, A. Can, M. Lagrange, and J.-F. Petiot. Creation of a corpus of realistic urban sound scenes with controlled acoustic properties. *The Journal of the Acoustical Society of America*, 141(5):4044–4044, May 2017.

[49] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.