Road traffic sound level estimation from realistic urban sound mixtures by Non-negative Matrix Factorization

Jean-Rémy Gloaguen, Arnaud Can, Mathieu Lagrange, Jean-François Petiot

Road traffic sound level estimation from realistic urban sound mixtures by Non-negative Matrix Factorization

Jean-Rémy Gloaguen[a,*], Arnaud Can[a], Mathieu Lagrange[b], Jean-François Petiot[b]

[a] *Ifsttar Centre de Nantes, UMRAE, Allée des Ponts et Chaussés, 44344 Bouguenais, France*
[b] *LS2N, 1 rue de Noë, 44331 Nantes, France*

Abstract

Experimental acoustic sensor networks are currently tested in large cities, and appear more and more as a useful tool to enrich modeled road traffic noise maps through data assimilation techniques. One challenge is to be able to isolate from the measured sound mixtures acoustic quantities of interest such as the sound level of road traffic. This task is anything but trivial because of the multiple sound sources that overlap within urban sound mixtures.

In this paper, the Non-negative Matrix Factorization (NMF) framework is developed to estimate road traffic noise levels within urban sound scenes. To evaluate the performances of the proposed approach, a synthetic corpus of sound scenes is designed, to cover most common soundscape settings, and whom realism is validated through a perceptual test. The simulated scenes reproduce then the sensor network outputs, in which the actual occurrence and sound level of each source are known.

Several variants of NMF are tested. The proposed approach, named threshold initialized NMF, appears to be the most reliable approach, allowing road traffic noise level estimation with average errors of less than 1.3 dB over the tested corpus of sound scenes.

Keywords: Non-negative Matrix Factorization, urban sound environment, road traffic sound level estimation

1. Introduction

In response to the growing demand from urban dwellers for a better environment, noise mapping has been recommended as a tool to tackle noise pollution. The enactment of the European Directive 2002/EC/49 makes such maps mandatory to cities over 100 000 inhabitants. Those maps play an important informative role, establishing the distribution of the sound levels all over the cities as well as the estimation of the number of city dwellers exposed to high sound level (> 55 dB(A)) [1]. Road traffic concentrates particular attention as it is the main urban source of noise annoyance. Road traffic noise maps are typically built from data collection that consist of traffic data collected on the main roads (flow rates, mean speeds and heavy vehicle ratio) and urban geographic data (building heights and location, topology, ground surfaces ...). Follows sound emission and sound propagation computational techniques, resulting in the production of the two indicators equivalent A-weighted sound levels, L_{DEN} (*Day-Evening-Night*) and L_N (*Night*) [2]. This procedure also enables drawing up action plans to reduce the noise exposure. Despite their unanimously recognized interest, noise maps suffer from some limitations. The computing cost required to produce noise maps at the city scale calls simplifications of the numerical tools and the simulation models that both generate uncertainties [3, 4]. Data collection is

*Corresponding author
Email address: jean-remy.gloaguen@ifsttar.fr (Jean-Rémy Gloaguen)

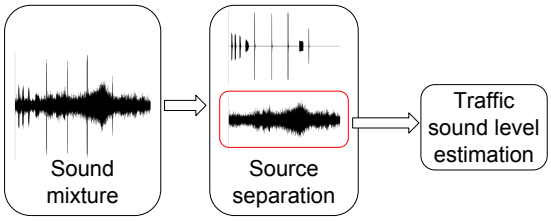


Figure 1: Block diagram of the blind source separation model

itself also a vector of uncertainty. Moreover, the produced aggregated indicators do not model the sound levels evolution due to the traffic variations throughout the day.

Noise measurements are thus increasingly used in addition to simulation to describe urban noise environments [5–7]. Several measurement set-ups have been proposed in the last years, including mobile measurements with high quality microphones [8, 9], participative sensing through dedicated smartphone applications [10, 11], or the development of fixed-sensor networks. In this latter case, the sensor networks can be based either on high-quality sensors as in [12, 13], or low-cost sensors as in the DYNAMAP project [14] or the CENSE project [15]. The costs and benefits of each protocol are discussed. Mobile and participatory measures increase spatial coverage at low cost, but lack temporal representativeness. Fixed networks are very reliable for measuring sound levels temporal variations, but allow only a small spatial coverage of the network. In addition, the low-cost sensors enable a wider deployment, but at the cost of increased uncertainties, the most extreme example being smartphone applications.

All these measurement protocols allow the combination of measures and predictions to improve the accuracy of the produced noise maps. Traffic noise maps and measurements were compared on restrictive areas in [16] and [17]. Wei et al. [18] modify the acoustical parameters of the simulation thanks to noise measurements, while Mallet et al. [19] call for data assimilation techniques between models and measurements to reduce the uncertainty of the produced noise maps. However, these works make the implicit assumption that the noise measurements consist mainly of road traffic. In the aim to improve road traffic noise maps, the use of measurements has first to deal with the challenge to estimate correctly the road traffic sound level. Even if road traffic is predominant on many urban areas, urban sound environments are composed of many different overlapping sound sources (passing cars, voices, footsteps, car horn, whistling birds ...), what makes the task of estimating correctly the traffic sound level within an urban sound mixture not trivial.

Many works have dealt with the classification [20, 21], the detection [22, 23] or the recognition [24, 25] of urban sound events. In these cases, a two-step scheme is followed where audio samples are described with a set of features (Mel Frequency Cepstral Coefficient, MPEG-7 descriptors ...) and classified with the help of a classifier (Gaussian Mixtures Models, Artificial Neural Network ...) [26, 27]. The classifier is learnt from a learning database and is next applied on a test database to validate the algorithms. Dedicated to the traffic, in [28], an Anomalous Event Detection, based on MFCC features, is proposed with the specific aim to improve the traffic sound estimation. It is based on the detection of unwanted sound events in order to discard them.

An other approach, followed in this paper, is to consider the blind source separation paradigm which consists in the extraction of a specific signal inside a set of mixed signals, see Figure 1. From the different existing methods, Non-negative Matrix Factorization (NMF) [29], appears to be a relevant method for monophonic sensor networks. Many applications can be found for musical [30, 31] and speech [32, 33] contents. Dedicated to sound separation with environmental sounds, Immani and Kasaï [34] used NMF in a two steps sound separation with the help of time variant gain features. Dedicated to the traffic sound separation, a first study [35] has been conducted, in which diverse NMF estimation rules are compared, namely the supervised, the semi-supervised, and the threshold initialized NMF, have been applied on a large set of simulated sound

scenes. This corpus mixes 6 sound categories (*alert, animals, climate, humans, mechanics, transportation*) with a traffic component calibrated to different sound levels, according to the other sound classes (in the rest of the document, these sound classes, not related to the traffic component, are resumed as the *interfering* sound class), to obtain variable traffic predominance. The diversity of this corpus was made to assess the performances and the limits of each NMF formula. However, if this study reveals the interest of NMF for urban sound environments, the assessment of its performance on a corpus of realistic sound scenes must be carried out in order to implement it on a sensor network. Design urban sound mixtures makes it possible to access to many acoustic properties as the onset and offset time and the sound level of each sound class and especially the traffic component. The realistic aspect of such a corpus is essential to obtain sound scenes similar to recordings and to validate NMF performances. However, like all simulated process, the realism of the scenes must be perceptually verified.

In this paper, an urban sound corpus based on annotated urban recordings, and whose degree of realism is assessed through a perceptual test, is designed in order to estimate the traffic sound level with the help of the NMF framework. The different NMF approaches are described in section 2. Next, the corpus of urban sound scenes is presented in section 3, from the sound database built-up to its validation. The experimental protocol and the results are then presented and discussed in section 4 and 5.

2. Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is a linear approximation method proposed by Paatero and Tapper [36] and popularized by Lee and Seung [29]. It consists in approximating a non negative matrix $\mathbf{V} \in \mathbf{R}^{+}_{F\times N}$ by the product of two non negative matrices: $\mathbf{W}$, called *dictionary* (or basis), and $\mathbf{H}$, called the *matrix activation* with dimensions $F \times K$ and $K \times N$ respectively,

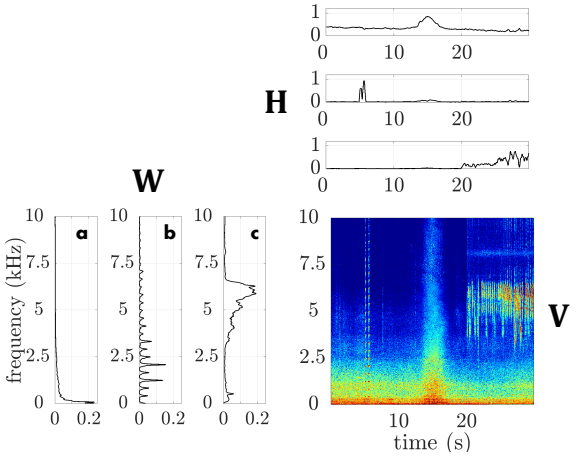


Figure 2: NMF decomposition of an audio spectrogram V composed of 3 elements ($K = 3$): passing car (a), car horn (b) and whistling bird (c).

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

The choice of the dimensions is often made such as $F \times K + K \times N < F \times N$ so that NMF can be a low rank approximation. This condition however is not mandatory. When applying NMF to audio data, $\mathbf{V}$ is usually considered as the magnitude spectrogram obtained by a Short-Time Fourier Transform, $\mathbf{W}$ includes audio spectra and $\mathbf{H}$ is equivalent to the temporal activation of each spectrum, see Figure 2. Because of the non-negativity constraint, only additive combinations between the elements of $\mathbf{W}$ are considered.

The approximation of $\mathbf{V}$ by $\mathbf{WH}$ product is defined by a cost function to minimize,

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

where $D(\bullet\|\bullet)$ is a divergence calculation such as:

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

$d_\beta(x|y)$ is usually chosen as a β-divergence [37], a subclasses belonging to the Bregman divergences [38] which include 3 specific divergence calculations: the Euclidean distance (eq. 4a), the Kullback-Leibler divergence (eq. 4b) and the Itakura-Saïto divergence (eq. 4c):

$$d_\beta(x|y) = \begin{cases} \frac{1}{2}(x-y)^2, & \beta = 2, \quad (4a) \\ x\log\frac{x}{y} - x + y, & \beta = 1, \quad (4b) \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0. \quad (4c) \end{cases}$$

The minimization problem (2) is solved iteratively by updating the form of matrices $\mathbf{W}$ and $\mathbf{H}$. Different algorithms such as Alternating Least Square Method [39] or Projected Gradient [40] have been considered. The most commonly used algorithm is the Multiplicative Update [41]. The latter method is chosen here, as it ensures non-negative results and the convergence of the results [37].

2.1. Supervised NMF

The most easiest case of NMF is the one where the sound sources can be known *a priori* and $\mathbf{W}$ can be built directly from audio samples. It leads to *supervised* NMF (SUP-NMF). $\mathbf{H}$ is then the only matrix to estimate and is updated at every iteration (eq. 5) [37]:

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

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 $\beta > 2$. Thus, the choice of the β-divergence in the equation 4 affects how the matrix $\mathbf{H}$ is updated. The $A \otimes B$ and A/B operators represent the Hadamard product and ratio.

Here, in an urban context, if the sound sources are known, their audio samples can be obtained to learn $\mathbf{W}$, see section 4.2. As the position of each element is indexed, the traffic source separation from the other sound sources is made by extracting, from the dictionary and the activation matrix, the related elements:

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

2.2. Semi-supervised NMF

The main issue with the supervised approach is the representational limit imposed by a fixed $\mathbf{W}$. To be completely successful, all the acoustical sources must be considered in the basis $\mathbf{W}$ which is not always possible in a complex urban environment. To overcome this issue, semi-supervised NMF (SEM-NMF) [42] has been proposed. It consists in decomposing, $\mathbf{W}_{F\times(K+J)}$ into two distinctive matrices: $\mathbf{W} = [\mathbf{W_s}\ \mathbf{W_r}]$ where $\mathbf{W_s}_{F\times K}$ is a fixed part of $\mathbf{W}$ composed of known audio spectra and $\mathbf{W_r}_{F\times J}$, a mobile part which is updated, see eq. 8a. Thus it is possible to let the method define the best elements to include in $\mathbf{W_r}$. Its dimension is set up as $J << K$ in order to consider, as a priority, the sound sources present in $\mathbf{W_s}$. $\mathbf{H}$ is then also decomposed in two matrices, $\mathbf{H}_{(K+J)\times N} = \begin{bmatrix} \mathbf{H_s} \\ \mathbf{H_r} \end{bmatrix}$. Eq. 1 becomes

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} = \mathbf{W_s}\mathbf{H_s} + \mathbf{W_r}\mathbf{H_r}. \quad (7)$$

$\mathbf{H_r}$ and $\mathbf{H_s}$ are updated separately, see eq. 8b to eq. 8c.

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

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

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

In this study, $\mathbf{W_s}$ is composed of traffic audio spectra, as it is the sound source of interest. Sources included in $\mathbf{W_r}$ are other sound sources (corresponding to the interfering class) that can be present in the urban sound scenes. The traffic signal estimation is next defined by the fixed part,

$$\tilde{\mathbf{V}}_{traffic} = [\mathbf{W_s}\mathbf{H_s}]. \quad (9)$$

The addition of $\mathbf{W_r}$ gives more flexibility to the method to represent correctly the spectrogram $\mathbf{V}$. The representational capability is increased, thus the approach is more adaptive to the different urban sound environments. Applications of SEM-NMF can be found for musical [43, 44] and speech contents [33, 45].

2.3. Thresholded Initialized NMF

To allow even more flexibility while still considering prior knowledge of the source of interest, we propose a third approach based on the unsupervised NMF framework: Threshold Initialized NMF (TI-NMF). Usually, in unsupervised NMF, the dictionary is initiated randomly when there is no *prior* knowledge on the sound sources present. Here, as the target sound source is known and the spectra are available, an initial dictionary, $\mathbf{W_0}$, is designed and then updated alternatively with $\mathbf{H}$,

$$\mathbf{W}^{(i+1)} \leftarrow \mathbf{W}^{(i)} \otimes \left(\frac{\left[\left(\mathbf{W}^{(i)} \mathbf{H} \right)^{(\beta-2)} \otimes \mathbf{V} \right] \mathbf{H}^T}{\left[\mathbf{W}^{(i)} \mathbf{H} \right]^{(\beta-1)} \mathbf{H}^T} \right)^{\gamma(\beta)}. \tag{10}$$

With this operation, $\mathbf{W_0}$ is oriented to the focused sound source (the road traffic) but also can be adapted to the content of the scene thanks to the updates. After N iterations, each element k of the final dictionary, $\mathbf{W'}$, is compared with its initial value in $\mathbf{W_0}$, in order to identify which element is stayed close to the traffic component. A cosine similarity $D_\theta\left(\mathbf{W_0} \| \mathbf{W'}\right)$ is computed for each element k as it is scale-invariant and bounded,

$$D_\theta\left(\mathbf{w_0} \| \mathbf{w'}\right) = \frac{\mathbf{w_0}.\mathbf{w'}}{\|\mathbf{w_0}\|.\|\mathbf{w'}\|}. \tag{11}$$

where $\mathbf{w}$ is a k element of $\mathbf{W}$ of $F \times 1$ dimensions. When $D_\theta\left(\mathbf{w_0} \| \mathbf{w'}\right)$=1, the element $\mathbf{w'}$ is identical to $\mathbf{w_0}$. If $D_\theta\left(\mathbf{w_0} \| \mathbf{w'}\right)$=1, both elements are fully different. The extraction of traffic elements in $\mathbf{W'}$ is carried out by a hard thresholding method [46]. It consists in weighting in a binary way $\mathbf{W'}$ according to $D_\theta\left(\mathbf{w_0} \| \mathbf{w'}\right)$ and a threshold value α_k such as:

$$\mathbf{w}_{traffic} = \alpha_k \mathbf{w'}. \tag{12}$$

with

$$\alpha_k = \begin{cases} 1 & \text{iff} \quad D_\theta\left(\mathbf{w}_0 \| \mathbf{w'}\right) > t_h, \quad (13a) \\ 0 & \text{else.} \quad (13b) \end{cases}$$

To summarize, to approximate the audio spectrogram $\mathbf{V}$ and estimate the traffic component, 3 methods are used which deal differently with the *a priori* knowledge on the traffic. SUP-NMF is only based on fixed *traffic* elements and is constrained to used it to obtain $\tilde{\mathbf{V}}_{traffic}$. SEM-NMF completes this dictionary by the add of a mobile dictionary to better consider the interfering sound events and then offer more flexibility. Finally, TI-NMF considers first a dictionary composed of *traffic* spectra, as SUP-NMF, but allows an update of them. The interest of an initiated dictionary is then to focus on the source of interest during the updates. This method therefore makes it more suitable for solving the generalization issues as it is the entire dictionary that can be fully adapted to the sound scenes. The elements that deviates too much from the originals spectra are then discarded by the thresholding step.

These above described methods are applied on an evaluation corpus, composed of simulated sound scenes, in order to compare the estimated traffic sound levels with an exact reference. This new sound corpus is designed by considering realistic urban sound environments with many kinds of mixed sound sources.

3. Design of realistic urban sound scenes

The evaluation corpus must be both realistic and non ambiguous in terms of traffic sound level. The former calls for the use of real urban sound environments, and the latter imposes the use of controlled sound sources with known

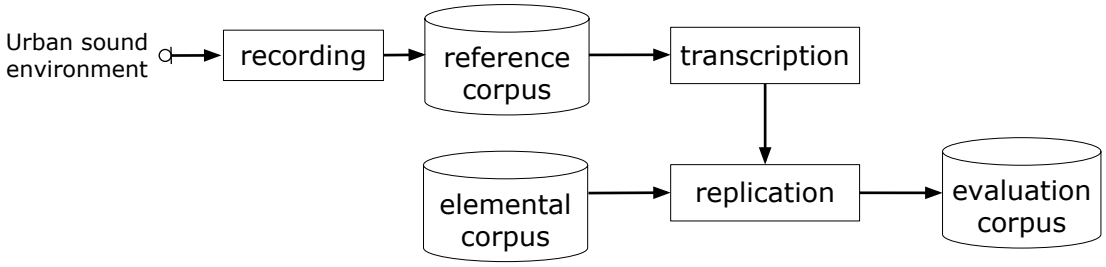


Figure 3: Bloc diagram of the design of the evaluation corpus

properties: onset and offset time location, and sound level. Indeed, the precise annotation of the sound level of a given source of interest in a real recording is a very complex problem. Two corpora are proposed to design a valid corpus of evaluation. The first, called the reference corpus, is a set of recordings of urban sound environments, from which, what we call, a transcription is made, *i.e.* which sound source is present from this time to this time. The second, called the elemental corpus, is a set of monophonic recordings of isolated sound sources that represent categories of events that are present in the reference scenes. The elemental corpus is built with still existing sounds and with specific recordings, see part 3.1. The evaluation corpus is then designed by replicating the reference corpus according to the transcription with the help of the elemental corpus, see Figure 3.

The reference corpus is composed of 76 recordings from 2 to 5 min, achieved in the 13th district of Paris (France) at 19 different locations [1], see Figure 4, which cover various sound environments. A complete description of the experimental protocol can be found in [47]. Two of the 76 recordings are rejected for the analysis because the audio files were corrupted, resulting in 74 valid audio files assumed as representative of the variety of urban sound environments. The recordings are listened and categorized within four different sound environments, as proposed in [48]: *park* (8 audio files with a cumulative duration of 16min01), *quiet street* (35 audio files with a cumulative duration of 77min27), *noisy street* (23 audio files with a cumulative duration of 56min10) and *very noisy street* (8 audio files with a cumulative duration of 21min42). Then, each audio file is transcripted, noticing the start, end time and level of each sound event along with its sound class. This annotation phase makes it possible to produce simulated sound scenes with the same positions of sound events than the recordings and therefore as close as possible to the real scenes.

Figure 4: Walked path with the 19 stop points [47].

3.1. Generation of the evaluation corpus

The sound scenes are generated with the *SimScene* software[2], [49], a simulation software generating monaural sound mixtures in wav format from an isolated sounds database. This software has already been used in a wide range of experiments for sound detection algorithm assessment [50, 51]. The *SimScene* software allows the design of

[1] Recordings were made as part of the Grafic project funded by Ademe

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

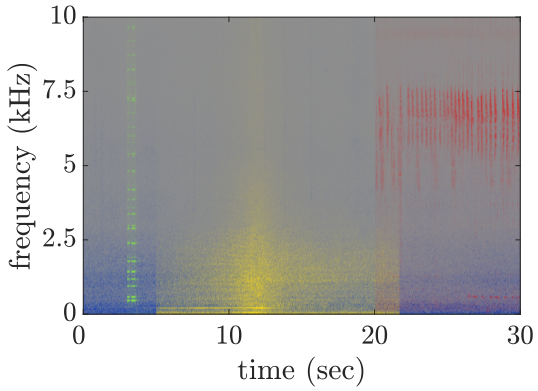


Figure 5: Spectrogram of a simple scene created with the *SimScene* software with one sound background (road traffic in blue) and 3 sound events (car horn in green, passing car in yellow and whistling bird in red).

several type of sequencing, from *abstract* ones (time indexes and amplitudes are drawn from random distributions) to precise ones, where the time indexes and amplitudes for each event are set by the user. The latter type is considered in this study. As output, *SimScene* generates an audio file of the global sound mixture as well as for each sound class present in the scene, see Figure 5. It makes it possible to know their exact contributions in the scene, in particular their sound levels.

To replicate the 74 recordings in simulated scenes, a high quality (wav format, 44.1 kHz sampling rate, high *Signal Noise Ratio*) elemental corpus has been built-up from audio samples found online (*freesound.org*) or with the help of an already existing sound database [52]. The elemental corpus is composed of two categories of sound: the *event* category, which includes 245 brief sound samples considered as salient, with a 1 to 20 seconds duration and classified among 21 sound classes (*ringing bell, whistling bird, car horn, passing car, hammer, barking dog, siren, footstep, metallic noise, voice ...*) and the *background* (or *texture*) category gathering 154 long duration sounds ($\approx$ 1mn30), whose acoustic properties do not vary in time. This category includes among others *whistling bird, crowd noise, rain, children playing in schoolyard, constant traffic noise* sound classes. Each sound class is composed of multiple samples (*carHorn01.wav, carHorn02.wav ...*) that are randomly chosen by the *simScene* software to bring diversity. As the road traffic is the main component in urban environment and is the sound source of interest, recordings of car passages have been made on the Ifsttar's runway. The recordings have been made for 4 different cars (Renault Scenic, Renault Mégane, Renault Clio and Dacia Sandero), at different speeds and gear ratios. Overall, 103 car passages have been recorded. In order to avoid overfitting issues, the audio samples of the first two cars (Renault Scenic and Renault Mégane) are included in the *SimScene*'s elemental corpus (50 audio files in total). The last 53 audio samples are dedicated to the dictionary design, $\mathbf{W}$ as part of NMF, see section 4.2. A full description of the recordings can be found in [53].

Table 1: Mean *Traffic Interfering Ratio* and its standard deviation for each sound environment.

	$mTIR$ (dB)
park	-9.10 ($\pm$ 7.35)
quiet street	0.88 ($\pm$ 5.92)
noisy street	6.96 ($\pm$ 5.16)
very noisy street	15.75 ($\pm$ 9.78)

With this built-up corpus, the *SimScene* software and the transcriptions of the recordings, 74 simulated sound scenes are generated, which have the same temporal structure of the reference recordings. The sound level of each sound class is adjusted manually on each sound scene to be faithful compared to the recorded scenes. To check this adjustment, the mean *Traffic Interfering Ratio* ($mTIR$) is calculated, see Table 1. It expresses, on all the scenes of a sound environment, the mean difference between the equivalent traffic sound levels of each scene, $L_{p,traffic}$, with the sound level of the *interfering* sound class, $L_{p,interfering}$, which gathered all the other sound sources not related to the traffic, see Eq. 14. It quantifies the predominance of the traffic component for the 4 types of sound environments,

$$mTIR = \frac{\sum_{i=1}^{M} L_{p,traffic} - L_{p,interfering}}{M}. \quad (14)$$

where M is the number of available scenes for each sound environment ($M = 8$ for *park* environment, $M = 35$ for *quiet street* environment, $M = 23$ for *noisy street* environment and $M = 8$ for *very noisy street* environment). The $mTIR$ is always negative for the *park* as it is the sound environment where the traffic is less present. The *interfering* sound class is therefore the main sound source which is coherent with this kind of environment. The $mTIR$ can be positive or negative depending on the traffic presence on the scene in the case of the *quiet street*. For the 2 others sound environments, when the traffic becomes the main sound source, $mTIR$ is always positive.

3.2. Perceptual test

To evaluate the level of realism of the evaluation corpus composed of replicated urban sound scenes, a perceptual test is considered.

The perceptual test is conducted with a panel of 50 listeners that are asked to assess the level of realism on a 7-point scale (1 is *not realistic at all*, 7 is *very realistic*) of a set of replicated and recorded scenes. The total number of sound scenes tested is set at 40. This is less than the number of scenes available to ensure that each audio sample is sufficiently assess. The first half includes 20 30-seconds audio files corresponding to the real scenes, including 5 scenes that belong to the sound environment *park*, 6 from *quiet street*, 4 from *noisy street* and 5 from *very noisy street* chosen randomly among the recorded scenes. The second half is composed of the replicated scenes from the evaluation corpus corresponding to the same 20 reference scenes. In order to limit the duration of the test to preserve the concentration of the subjects, each subject listens to only a subset of 20 sound scenes. All the scenes are normalized to the same sound level, chosen at 65 dB, and the subjects are not allowed to change the output sound level once set at the beginning of the experiment.

The experimental design is elaborated following a partially Balanced Incomplete Block Design (PBIBD) [54] the audio allocated to each participant and the listening order. This process allows to each sound sample to be assessed almost the same number of time and to avoid statistical biais. The experimental design and the listening order per participant are performed with the package *sensoMineR* on the *R* software [55].

The test was available online from February 8th, 2017 and the needed number of participant (50) has been reached 12 days later. During the test, the participants had the possibility to listen to each scene as many times as wanted before assessing, without being able to change their judgment afterwards. The participants could also leave a comment on each audio to explain the rating [3]. Based on the information provided, the panel of 50 listeners was made of 31 males and 18 females (one not documented) with an average age of 36 ($\pm$ 12) years old. 62% of the participants declared having no experience in the listening of urban sound mixtures. The results show that the average score (with its standard deviation) of all the replicated scenes ($m_{replicated} = 5.1$ ($\pm$ 1.6)) is close to the recorded ones ($m_{recorded} = 4.9$ ($\pm$ 1.6)). To determine whether a replicated scene is perceived in a similar way to that recorded, a Student's t-test is performed, for each scene, between the scores from the recorded sample and those from the replicated sample with the H_0 hypothesis which considers the similarity between the distribution of the rates for the recorded and the replicated scenes of each participant with a *p-value* threshold of 5%. The 20 performed t-tests show that the differences in the assessment of the replicated and the recorded scene, according to their degree or realism, is not significant (p-value$>$ 5%).

More details on the results can be found in [53]. As the perceived realism of the replicated and the recorded scenes

[3] The interface of the experiment is available `http://soundthings.org/research/xpRealism`

are not significantly different, we consider that these sound mixtures are relevant to assess the performances of NMF according to the traffic sound level estimate.

4. Methods for the performance evaluation

The experiment aims at evaluating in a meaningful manner the performance of the NMF approach. To do so, the 74 replicated sound scenes, organized on 4 different types of sound environments are iteratively fed to the estimator which output an estimate of the equivalent sound level of the traffic within each scene $\tilde{L}_{p,traffic}$ (dB). This result is then compared to the reference sound level value given by the simulation process, $L_{p,traffic}$, see Figure 6.

4.1. Reference estimator

A reference estimator is necessary to be able to compare the performances of the different NMF-methods. As the road traffic is mainly composed of a low frequency content, a frequency low-pass filter (LP filter) is considered as baseline. The estimation of the traffic sound level with this reference estimator is simply the sound level after low-pass filtering,

$$\tilde{\mathbf{V}}_{traffic} = \mathbf{V}_{f_c}. \tag{15}$$

Different cut-off frequencies have been chosen such as $f_c \in \{500,\ 1\text{k},\ 2\text{k},\ 5\text{k},\ 10\text{k},\ 20\text{k}\}$ Hz. The experimental factors related to this estimator are summarized in Table 2.

The second estimator is based on the three NMF formula presented in part 2 (see Figure 8). Multiple experimental factors are involved in this second estimator where each of them having different modalities.

4.2. Dictionary building for NMF-methods

The dictionary is designed from a sound database specially dedicated to this task. To prevent any overfitting issues, it contains the 53 audio files of the 2 cars not included in the creation of the evaluation corpus, see part 3.1.

First the spectrogram of each audio file is computed ($w = 2^{12}$ sample points with 50 % overlap). The spectrogram is then cut in multiple temporal frames of $\mathbf{w_t} \in \{0.5,\ 1\}$ second duration. In each frame, the root mean square on each frequency bin is calculated to obtained a spectrum of $F \times 1$ dimension. This method allows the description of the audio sample with a finer spectra and then having the different characteristic pitches of the traffic spectra. An illustrative example on a 3 seconds sample is displayed in Figure 7. From the 53 audio files, we obtain respectively 2218 elements for $\mathbf{w_t} = 0.5$ second and 1109 elements for $\mathbf{w_t} = 1$ second.

A $\mathbf{K}$-means clustering algorithm is applied to those elements to reduce these dimensions to $\mathbf{K} \in \{25,\ 50,\ 100,\ 200\}$ in order to avoid redundant information and decrease the computation time. The resulting $\mathbf{K}$ cluster centroids are taken as the elements of $\mathbf{W}$. Each basis of $\mathbf{W}$ is then normalized as $\|\mathbf{w}\| = 1$ where $\|\bullet\|$ is the ℓ_1 norm. Table 2 summarizes the different modalities of the two experimental factors ($\mathbf{K}$ and $\mathbf{w_t}$).

4.3. NMF experimental factors

NMF is performed for 3 β-divergences: $\beta = 2$ (Euclidean distance), $\beta = 1$ (Kullback-Leibler divergence) and $\beta = 0$ (Itakura-Saïto divergence). The spectrogram $\mathbf{V}$ and the dictionary $\mathbf{W}$ are expressed in a logarithmic scale through a third octave band representation that reduces the high frequency predominance where the traffic component is absent. In addition, as the number of frequency bins is reduced ($F = 29$), the computation time is reduced too. 400 iterations are performed to get a stabilized results. For SEM-NMF, the number of elements in $\mathbf{W_r}$ is set to $J = 2$. For hard thresholding, the threshold value, $\mathbf{t_h}$ is set between 0.30 and 0.60 with a 0.01 increment. Each unique association of modalities between each ex-

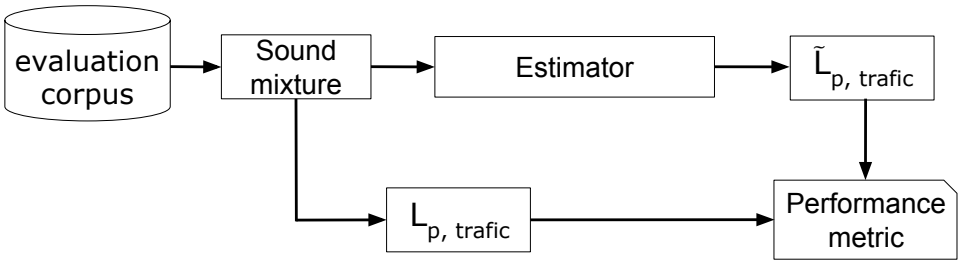


Figure 6: Bloc diagram of the different stages for the estimation of the traffic sound level.

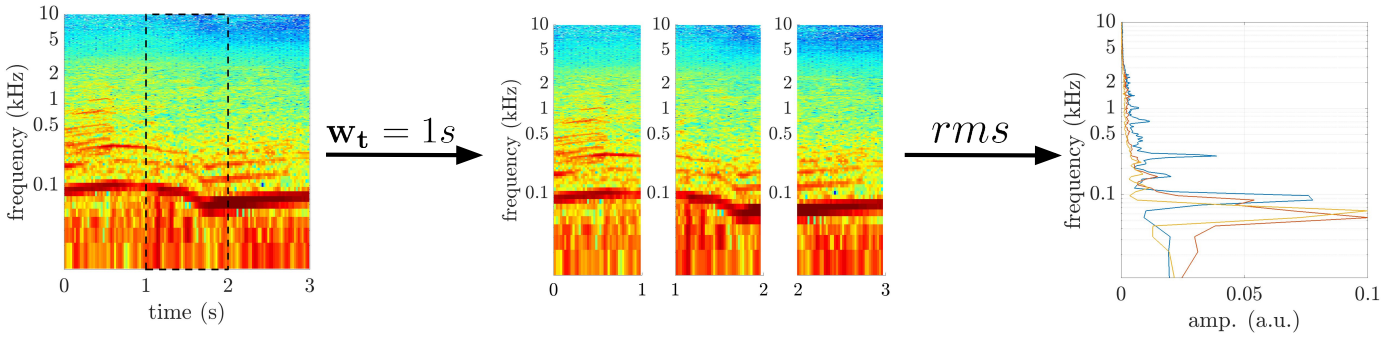


Figure 7: Dictionary building on a 3 second example of a passing car with $\mathbf{w_t} = 1$ second, in dashed line a temporal frame of 1 s. In this case, the dictionary $\mathbf{W}$ is made of 3 spectra, each representative of a texture frame of $\mathbf{w_t}$ duration.

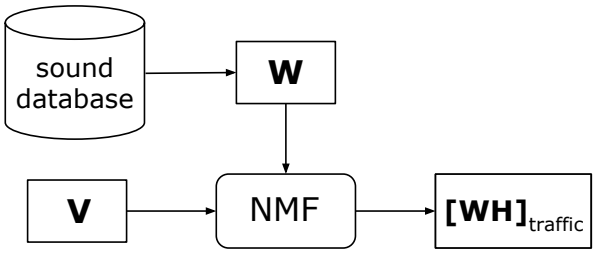


Figure 8: Bloc diagram of the NMF estimator.

perimental factor forms an experimental setting. For the filter estimator, 24 settings are computed (4×6). For SUP and SEM-NMF, 192 settings are computed ($4 \times 2 \times 2 \times 4 \times 3$). Finally, for TI-NMF, the number of settings is much higher (2976) due to the higher cardinality of the set of threshold values ($4 \times 1 \times 2 \times 4 \times 3 \times 31$). The experimental factors and their different modalities are displayed in Table 2.

The approximated traffic spectrograms $\tilde{\mathbf{V}}_{traffic}$ are obtained after 400 iterations. The estimated traffic sound level in dB, $\tilde{L}_{p,traffic}$, is then computed,

$$\tilde{L}_{p,traffic} = 20 \log \left(\frac{p_{rms}}{p_0} \right), \qquad (16)$$

where p_0 is the reference sound pressure, $p_0 = 2 \times 10^{-5}$ Pa. For each setting, M traffic sound levels, corresponding to the M scenes of each sound environment, are then calculated.

4.4. Metrics

The traffic sound levels, $\tilde{L}_{p,traffic}$, are compared to the exact values, $L_{p,traffic}$, through the Mean Absolute Error (MAE) [56]. The MAE consists in the average across of the absolute difference between the exact and the estimated sound levels,

$$MAE_j = \left[\frac{\sum_{i=1}^{M} |L^i_{p,traffic} - \tilde{L}^i_{p,traffic}|}{M} \right]_j, \qquad (17)$$

for each setting j. But it is also possible to average this metric according to the 4 different sound environments, through the mean MAE error $mMAE$, to estimate the

Table 2: Experimental factors and their modalities for the NMF estimator.

experimental factors	**modalities**						**number of modalities**
sound environment	park 'P'		quiet street 'Q'	noisy street 'N'		very noisy street 'vN'	4
method	LP Filter		SUP NMF	SEM NMF		TI NMF	4
$\mathbf{f_c}$ (kHz)	0.5	1	2	5	10	20	6
$\mathbf{w_t}$ (s)		0.5			1		2
K	25		50	100		200	4
β	0			1		2	3
hard threshold $\mathbf{t_h}$	from 0.30 to 0.60 with a 0.01 step						31

optimal setting that offers the lowest error for all the sound environments:

$$mMAE = \frac{\sum_{i=1}^{4} MAE_i}{4}, \quad (18)$$

where the other experimental factors (estimator, f_c, $\mathbf{K}$, $\mathbf{w_t}$, β, threshold value $\mathbf{t_h}$) are fixed.

5. Results and discussion

Table 3 summarizes the lowest $mMAE$ errors according to the estimator (LP filter, NMF) and β with the best setting of the experimental factors.

The LP filter with f_c = 20 kHz cut-off frequency is equivalent to consider the sound level of the entire scene without specific distinction between the sound sources. The error is then important with a high standard deviation ($mMAE$ = 3.76 (± 4.35) dB). The lowest error for a LP filter is obtained with f_c = 500 Hz ($mMAE$ = 2.14 (± 1.83) dB).

When considering all the sound scenes, SUP-NMF does not succeed to achieve a lower error than the 500 Hz LP filter for all the β values. By adding the mobile part $\mathbf{W_r}$ in the dictionary, SEM-NMF with $\beta = 0$ and $\beta = 1$ allows a lower error than 500 Hz LP filter with a reduced standard deviation especially for $\beta = 1$ ($mMAE$ = 1.94 (± 0.38) dB).

TI-NMF is the approach with the lowest global error (< 1.50 dB). The best result is obtained for TI-NMF (MAE = 1.24 (± 1.24) dB) with $\beta = 2$, $\mathbf{K} = 200$, $\mathbf{w_t} = 0.5$ s and as threshold value $\mathbf{t_h} = 0.32$. This combination of settings offers the most efficient method adapted to the different sound environments. Furthermore, on the dictionary creation, only SEM-NMF proposes the same dictionary design for all the best methods according to β. In the opposite, SUP and TI-NMF propose different associations between $\mathbf{K}$ and $\mathbf{w_t}$. One can notice that, through the 3 methods, it is mainly a high number of elements in $\mathbf{K}$ (100, 200) that is preferred. With more elements, it makes it possible to resolve easily the generalization issue.

From these global results, the MAE errors are compared to the LP filter and each method for the 4 types of sound environments, see Figure 9.

Aside SEM-NMF, all the methods show the same error evolution: a decrease of the error with the increase of the traffic predominance. On contrary, SEM-NMF shows an almost constant error for all 4 sound environments. The LP filter error is mainly important for environments where the traffic is less present. As this approach considers the

Table 3: Best $mMAE$ errors according to the experimental factor β and the traffic sound level assessment method (in bold letter, the lowest error).

method	f_c (kHz)	β	**K**	$\mathbf{w_t}$ (s)	$\mathbf{t_h}$	$mMAE$ (dB)
filter	20	-	-	-	-	3.76 ($\pm$ 4.35)
filter	0.5	-	-	-	-	2.14 ($\pm$ 1.83)
SUP-NMF	-	0	200	0.5	-	4.06 ($\pm$ 4.69)
SUP-NMF	-	1	200	0.5	-	2.79 ($\pm$ 3.38)
SUP-NMF	-	2	25	1	-	2.32 ($\pm$ 2.80)
SEM-NMF	-	0	200	1	-	2.05 ($\pm$ 0.70)
SEM-NMF	-	1	200	1	-	1.94 ($\pm$ 0.38)
SEM-NMF	-	2	200	1	-	2.39 ($\pm$ 1.23)
TI-NMF	-	0	25	1	0.39	1.42 ($\pm$ 0.89)
TI-NMF	-	1	100	1	0.35	1.38 ($\pm$ 0.88)
TI-NMF	-	**2**	**200**	**0.5**	**0.32**	**1.24 ($\pm$ 1.24)**

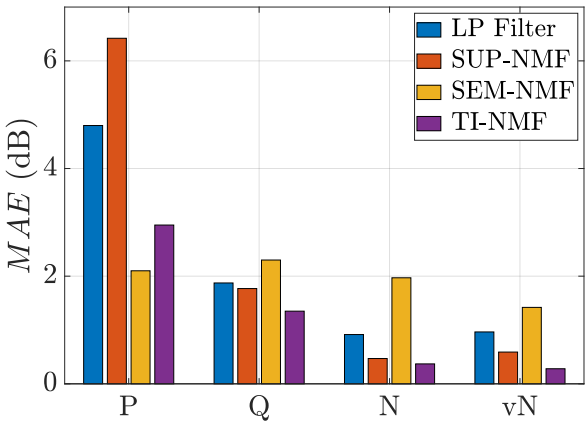
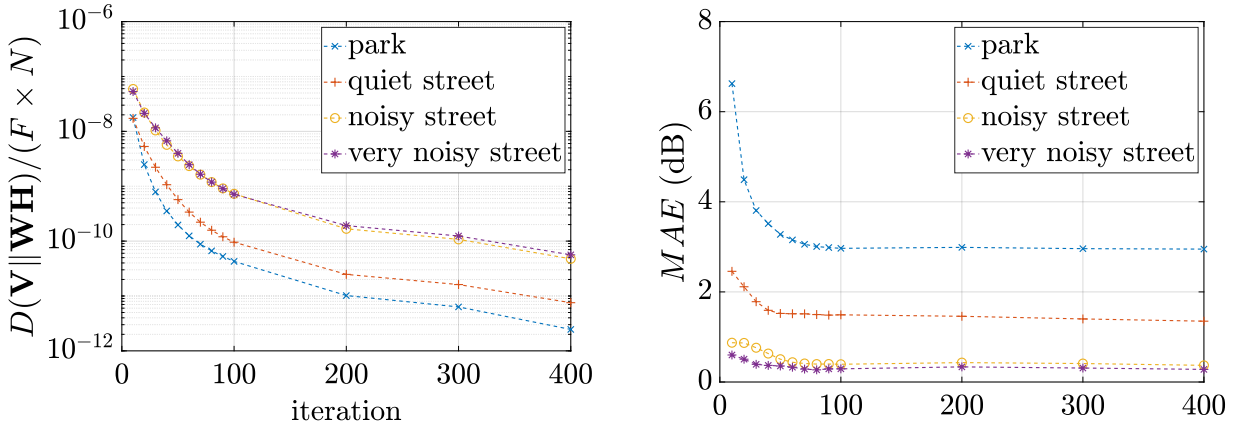


Figure 9: MAE errors with the standard deviations according to each sound environment for the best combination of the LP filter (f_c = 500 Hz), SUP-NMF ($\beta = 2$, $\mathbf{K} = 25$, $\mathbf{w_t} = 1$ second), SEM-NMF ($\beta = 1$, $\mathbf{K} = 200$, $\mathbf{w_t} = 1$ second) and TI-NMF ($\beta = 2$, $\mathbf{K} = 200$, $\mathbf{w_t} = 0.5$ second, $\mathbf{t_h} = 0.32$).

remaining energy as the traffic component, no distinction is made between the different sound sources not related to the traffic component. The interfering sound class is then wrongly considered as traffic component. On the opposite, for noisy and very noisy environments, the performances of the LP filter are good ($MAE < 1$ dB). The errors are here due to a high deletion of the traffic energy by the filter while it becomes the main sound source. Consequently, the 500 Hz LP filter estimator provides a low MAE error through the sound environments thank to a balance between the remaining and discarded energy.

Despite a fixed dictionary composed of traffic spectra, SUP-NMF fails to identify correctly the traffic component particularly for *park* ($MAE = 6.42$ dB) environments. With this method, as NMF minimizes the cost function, eq. 2, the dictionary's elements are used to model the other sound sources which can not allow a rightful approximation of the traffic component. On the opposite, for *noisy* and *very noisy* environments, SUP-NMF identifies correctly the traffic components ($MAE < 0.6$ dB) as it is the main source. In the case of SEM-NMF, adding the mobile dictionary, $\mathbf{W_r}$, makes it possible to include the other sound sources not present in the dictionary. If this behavior is advantageous for the *park* environment ($MAE = 2.10$ dB) where lot of different kind of sources are present, it is less advantageous for the rest of the environments where the traffic becomes predominant resulting



(a) Evolution of the cost function.

(b) Evolution of the MAE errors.

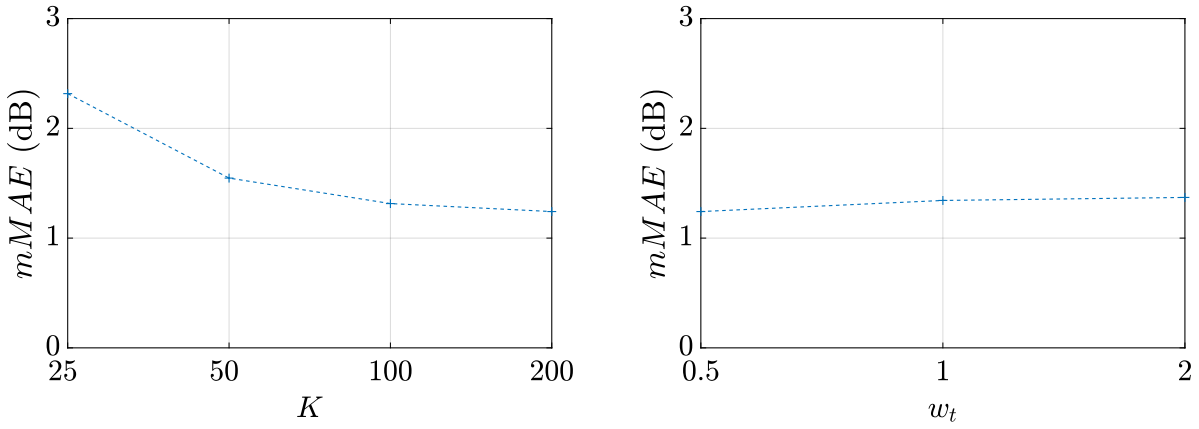
Figure 10: Normalized reconstruction error of the global sound mixtures a and the traffic component b for each sound environment.

in the highest errors. Indeed, this degree of freedom generates higher error as $\mathbf{W_r}$ is not constrained and is free to include traffic component in it, penalizing the traffic sound level estimation.

Finally, TI-NMF presents the most performing results, except in the *park* environment ($MAE = 2.95$ dB). In this sound environment, with the threshold value $t_h = 0.32$, the traffic dictionary is composed, on average, of 136 elements. By increasing this value, it would be possible to put aside the spectra furthest away of the traffic elements and then decrease the error. For the rest of the sound environments, TI-NMF has the lowest errors. For very noisy environment the error is even very low ($MAE = 0.28$ dB). With on average 198 elements in $\mathbf{W}_{traffic}$, almost all the elements present in $\mathbf{W}'$ can be considered as traffic elements. Considering a unique dictionary fitted to the sound scene under evaluation thus makes TI-NMF very effective when traffic is predominant, while the thresholding step makes it possible to discard the elements of the dictionary that deviate too much when the traffic is less present.

To better understand the behavior of TI-NMF, the evolution of the normalized cost function ($D(\mathbf{V}\|\mathbf{WH})/(F \times N$) as the duration of each audio sample is different) and the error MAE as the function of the number of iterations are displayed for each sound environment in Figure 10. For each sound environment, the cost function is very low after 400 iterations, meaning that $\mathbf{WH}$ approximates correctly $\mathbf{V}$. However, even if the convergence is not reached (Figure 10a), the traffic error reconstruction is constant after the 200th iteration (Figure 10b), which means that the updates of $\mathbf{W}$ and $\mathbf{H}$ are mostly dedicated to the interfering sound sources. Also, even if the cost function is higher for the *very noisy street* than the *park*, the traffic sound level estimation is better in the *very noisy street* sound environment due to the higher presence of this component in it.

Finally, the influence of the number of elements $\mathbf{K}$ in $\mathbf{W}$ and the size of the temporal frame $\mathbf{w_t}$ on the $mMAE$ error made with TI-NMF, with the other experimental factors keep constant, are summarized in Figure 11. Naturally, the traffic error reconstruction decreases, with the increase of the number of basis in $\mathbf{W}$, as it allows more basis to describe the traffic sound source. If this improvement is major between 25 and 100 elements, its influence decreases between 100 and 200 elements suggesting that the increase of the size of $\mathbf{K}$ will not improve significantly the quality of the estimation. The temporal frame $\mathbf{w_t}$ is a less influential parameter for TI-NMF. As the matrices are updates at each iteration, it makes sense that the variations in the spectra shape in $\mathbf{W_0}$, due to this experimental



(a) Influence of the experimental factor **K** (TI-NMF, w_t = 0.5 s, $\beta = 2$, $t_h = 0.32$, 400 iterations).

(b) Influence of the experimental factor $\mathbf{w_t}$ (TI-NMF, K = 200 s, $\beta = 2$, $t_h = 0.32$, 400 iterations).

Figure 11: Influence of the number of elements **K** (a) and the temporal frame $\mathbf{w_t}$ (b) on the $mMAE$ error with the best TI-NMF obtained in Table 3.

factor, will disappear after a set of iterations.

6. Conclusion

The non-negative matrix factorization framework has been considered as a source separation tool to estimate the traffic sound level from a corpus of urban sound scenes artificially built. Those scenes are designed to be as similar as possible to the outputs of a deployed sensor networks with the advantages of the simulation process (sound level and position of each source controlled and known). The realism of the scenes has been verified thanks to a perceptual test.

The results confirm the potential of the NMF method on such application scenario as it takes into account the overlap between the multiple sound sources present in cities and is suited to monophonic sensor networks. Different NMF algorithmic schemes have been studied through the supervised and semi-supervised approach. On all the sound environments, these common approaches reveal to be not sufficiently efficient: supervised NMF approach, with its fixed dictionary, does not succeed to estimate correctly the traffic sound level especially when this sound source is quiet, while semi-supervised approach with the presence of a mobile part in the dictionary is the best estimator for *park* environments but fails on heavily traffic scenes.

The proposed approach, named Thresholded Initialized NMF, achieved the lowest error in the evaluation corpus. Consequently, in the case where the location or the type of sound environments the sensors are monitoring cannot be identified (for instance within a mobile measurement framework), TI-NMF appears to be the most appropriate method. If the sound environment can be identified through a prior analysis, or based on positioning data [57, 58], it should be possible to adapt the estimation procedure by selecting the most efficient approach in order to further reduce the error in the estimated road traffic sound levels.

Further analyses are required to extend the proposed method to other sound sources, such as birds or voices sounds, which can conveniently be done by replacing or adding elements in the dictionary. This utilization would be useful in the context of muti-source noise mapping that is gaining interest [59, 60]. Finally, the parameters selected in this study are valid for this evaluation corpus. Further analyses on various corpus of sound scenes are needed to evaluate the robustness of the method and select the most

relevant approaches for specific sound environments (predominance of water or industrial sounds, rural environments...).

For reproducibility purposes, the evaluation corpus, the experimental protocol and the programs developed under the Matlab software are available online [4] as the sound dataset [5].

Acknowledgements

The authors would like to thank Pierre Aumond and Catherine Lavandier from the University of Cergy-Pontoise for transmitting us the data of the *Grafic* project.

Funding

This study is co-funded by Ifsttar and Pays de la Loire region.

References

[1] C. Nugent, N. Blanes, J. Fons, et al., Noise in europe 2014, European Enviroment Agency 10 (2014) 2014.

[2] S. Kephalopoulos, M. Paviotti, F. A. Ledee, Common noise assessment methods in europe (cnossos-eu), Common noise assessment methods in Europe (CNOSSOS-EU) (2012) 180.

[3] M. Arana, R. San Martín, I. Nagore, D. Pérez, What precision in the digital terrain model is required for noise mapping?, Applied Acoustics 72 (8) (2011) 522–526.

[4] H. Van Leeuwen, S. Van Banda, Noise mapping-state of the art-is it just as simple as it looks?, Proceedings of EuroNoise 2015.

[5] G. R. Gozalo, J. M. B. Morillas, J. T. Carmona, D. M. González, P. A. Moraga, V. G. Escobar, R. Vílchez-Gómez, J. A. M. Sierra, C. Prieto-Gajardo, Study on the relation between urban planning and noise level, Applied Acoustics 111 (2016) 143–147.

[6] P. H. T. Zannin, M. S. Engel, P. E. K. Fiedler, F. Bunn, Characterization of environmental noise based on noise measurements, noise mapping and interviews: A case study at a university campus in brazil, Cities 31 (2013) 317–327.

[7] A. Can, T. Van Renterghem, D. Botteldooren, Exploring the use of mobile sensors for noise and black carbon measurements in an urban environment, in: S. F. d'Acoustique (Ed.), Acoustics 2012, Nantes, France, 2012.

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

[9] A. Can, L. Dekoninck, D. Botteldooren, Measurement network for urban noise assessment: Comparison of mobile measurements and spatial interpolation approaches, Applied Acoustics 83 (2014) 32–39.

[10] J. Picaut, P. Aumond, A. Can, et al., Noise mapping based on participative measurements with a smartphone, in: Acoustics '17 Boston, Vol. 141 of The Journal of the Acoustical Society of America, Acoustical Society of America and the European Acoustics Association, Boston, United States, 2017, p. 3808.

[11] R. Ventura, V. Mallet, V. Issarny, et al., Evaluation and calibration of mobile phones for noise monitoring application, The Journal of the Acoustical Society of America 142 (5) (2017) 3084–3093.

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

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

[14] X. Sevillano, J. C. Socoró, F. Alías, et al., DYNAMAP–development of low cost sensors networks for real time noise mapping, Noise Mapping 3 (1) (2016) 172–189.

[15] J. Picaut, A. Can, J. Ardouin, 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) (2017) 3808–3808.

[16] N. Lefebvre, X. Chen, P. Beauseroy, M. Zhu, Traffic flow estimation using acoustic signal, Engineering Applications of Artificial Intelligence 64 (2017) 164–171.

[17] P. Mioduszewski, J. A. Ejsmont, J. Grabowski, D. Karpiński, Noise map validation by continuous noise monitoring, Applied Acoustics 72 (8) (2011) 582–589.

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

[19] R. Ventura, V. Mallet, V. Issarny, et al., Estimation of urban

[4] https://github.com/jean-remyGloaguen/articleNmfTrafficSimScene2018

[5] https://zenodo.org/record/1184443

noise with the assimilation of observations crowdsensed by the mobile application ambiciti, in: INTER-NOISE and NOISE-CON Congress and Conference Proceedings, Vol. 255, Institute of Noise Control Engineering, 2017, pp. 5444–5451.

[20] X. Valero, F. Alías, Hierarchical classification of environmental noise sources considering the acoustic signature of vehicle pass-bys, Archives of Acoustics 37 (4) (2012) 423–434.

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

[22] B. Luitel, Y. S. Murthy, S. G. Koolagudi, Sound event detection in urban soundscape using two-level classification, in: Distributed Computing, VLSI, Electrical Circuits and Robotics, IEEE, 2016, pp. 259–263.

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

[24] B. Defreville, F. Pachet, C. Rosin, P. Roy, Automatic Recognition of Urban Sound Sources, in: F. Paris (Ed.), AES 120th ConventionAudio Engineering Society, Audio Engineering Society, Paris, France, 2006, p. 9.

[25] G. Parascandolo, H. Huttunen, T. Virtanen, Recurrent neural networks for polyphonic sound event detection in real life recordings, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 6440–6444.

[26] S. Chu, S. Narayanan, C.-C. J. Kuo, Environmental sound recognition using mp-based features, in: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, 2008, pp. 1–4.

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

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

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

[30] P. Smaragdis, J. Brown, Non-negative matrix factorization for polyphonic music transcription, in: Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on., 2003, pp. 177–180.

[31] E. Benetos, M. Kotti, C. Kotropoulos, Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection, in: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, Vol. 5, IEEE, 2006, pp. V–V.

[32] K. W. Wilson, B. Raj, P. Smaragdis, A. Divakaran, Speech denoising using nonnegative matrix factorization with priors, in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 4029–4032.

[33] G. J. Mysore, P. Smaragdis, A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 17–20.

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

[35] J.-R. Gloaguen, M. Lagrange, A. Can, J.-F. Petiot, Estimation of road traffic sound levels in urban areas based on non-negative matrix factorization techniques, in revision (2018).

[36] P. Paatero, U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics 5 (2) (1994) 111–126.

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

[38] R. Hennequin, B. David, R. Badeau, Beta-Divergence as a Subclass of Bregman Divergence, IEEE Signal Processing Letters 18 (2) (2011) 83–86.

[39] A. Cichocki, 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, Springer, Berlin, Heidelberg, 2007, pp. 793–802.

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

[41] D. Lee, H. Seung, Algorithms for Non-negative Matrix Factorization, in: In NIPS, MIT Press, 2000, pp. 556–562.

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

[43] F. Weninger, J. Feliu, 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, IEEE, 2012, pp. 61–64.

[44] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, K. Kondo, Music Signal Separation Based on Supervised Nonnegative Matrix Factorization with Orthogonality and Maximum-Divergence Penalties, IEICE Transactions on Funda-

mentals of Electronics, Communications and Computer Sciences E97.A (5) (2014) 1113–1118.

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

[46] D. L. Donoho, I. M. Johnstone, Threshold selection for wavelet shrinkage of noisy data, in: Engineering in Medicine and Biology Society. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE, Vol. 1, IEEE, 1994, pp. A24–A25.

[47] P. Aumond, A. Can, B. De Coensel, et al., Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context, Acta Acustica united with Acustica 103 (3) (2017) 430–443.

[48] A. Can, B. Gauvreau, Describing and classifying urban sound environments with a relevant set of physical indicators, The Journal of the Acoustical Society of America 137 (1) (2015) 208–218.

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

[50] G. Lafay, M. Rossignol, N. Misdariis, et al., A New Experimental Approach for Urban Soundscape Characterization Based on Sound Manipulation : A Pilot Study, in: International Symposium on Musical Acoustics, Le Mans, France, 2014, pp. 593–599.

[51] E. Benetos, G. Lafay, M. Lagrange, M. D. Plumbley, Detection of overlapping acoustic events using a temporally-constrained probabilistic model, in: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE, 2016, pp. 6450–6454.

[52] J. Salamon, C. Jacoby, J. P. Bello, A dataset and taxonomy for urban sound research, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 1041–1044.

[53] J.-R. Gloaguen, A. Can, M. Lagrange, J.-F. Petiot, Creation of a corpus of realistic urban sound scenes with controlled acoustic properties, in: Acoustics '17 Boston, Vol. 141 of The Journal of the Acoustical Society of America, Acoustical Society of America and the European Acoustics Association, Boston, United States, 2017, pp. 4044–4044.

[54] J. John, T. J. Mitchell, Optimal incomplete block designs, Journal of the Royal Statistical Society. Series B (Methodological) (1977) 39–43.

[55] S. Lê, F. Husson, SensoMineR: A package for sensory data analysis, Journal of Sensory Studies (2008) 14 – 25.

[56] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, Climate research 30 (1) (2005) 79–82.

[57] A. Can, G. Guillaume, B. Gauvreau, Noise indicators to diagnose urban sound environments at multiple spatial scales, Acta Acustica united with Acustica 101 (5) (2015) 964–974.

[58] C. Lavandier, P. Aumond, S. Gomez, C. Dominguès, Urban soundscape maps modelled with geo-referenced data, Noise Mapping 3 (1) (2016) 278–294.

[59] P. Aumond, L. Jacquesson, A. Can, Probabilistic modeling framework for multisource sound mapping, Applied Acoustics 139 (2018) 34–43.

[60] F. Aletta, J. Kang, Noise indicators to diagnose urban sound environments at multiple spatial scales, Noise Mapping 2 (1) (2015) 1–12.