

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

Jean-Rémy Gloaguen<sup>a,\*</sup>, Arnaud Can<sup>a</sup>, Mathieu Lagrange<sup>b</sup>, Jean-François Petiot<sup>b</sup>

<sup>a</sup>*Ifsttar Centre de Nantes, UMRAE, Allée des Ponts et Chaussées, 44344 Bouguenais, France*

<sup>b</sup>*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, a Non-negative Matrix Factorization (NMF) framework is developed to estimate road traffic noise levels within urban sound scenes. To evaluate the performance of the proposed approach, a synthetic corpus of sound scenes is designed to cover most common soundscape settings. The realism of the produced scenes is validated through a perceptual test, forming a protocol that both reproduces the sensor network outputs, and in which the actual occurrence and sound level of each source is known.

Several variants of NMF are tested. The threshold initialized NMF appears to be the most reliable approach, allowing road traffic noise level estimation with average errors of 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, etc.). 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 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]. Data collection is itself also a vector of uncertainty. Moreover, the produced aggregated indicators do not model the sound levels evolution due to

---

\*Corresponding author

Email address: [jean-remy.gloaguen@ifsttar.fr](mailto:jean-remy.gloaguen@ifsttar.fr) (Jean-Rémy Gloaguen)

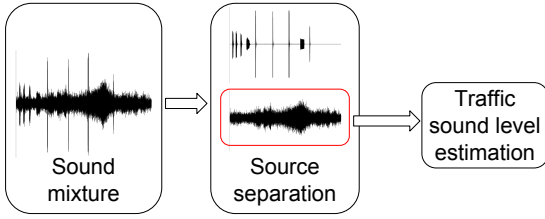


Figure 1: Block diagram of the blind source separation model

the traffic variations throughout the day.

Noise measurements are thus increasingly used in addition to simulation to describe urban noise environments. Several measurement set-ups have been proposed in the last years, including mobile measurements with high quality microphones [4, 5], participative sensing through dedicated smartphone applications [6, 7], or the development of fixed-sensor networks. In this latter case, the sensor networks can be based either on high-quality sensors as in [8, 9], or low-cost sensors as in the DYNAMAP project [10] or the CENSE project [11]. 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 allows 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 [12] and [13]. Wei et al. [14] modify the acoustical parameters of the simulation thanks to noise measurements, while Mallet et al. [15] 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 detection [16] or the recognition [17] of sound events in environmental sound scenes. 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 them with the help of a classifiers (Gaussian Mixtures Models, Artificial Neural Network ...) [18, 19]. The classifiers is learnt from a learning database and are next applied on a test database to validate the algorithms. Dedicated to the traffic, in [20], 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) [21], appears to be a relevant method for monophonic sensor networks. Many applications can be found for musical [22, 23] and speech [24, 25] contents. Dedicated to sound separation, Immani and Kasaï [26] used NMF in a two steps sound separation with the help of time variant gain features. A first study [27] 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 that which mixed traffic component with specific urban sounds at calibrated sound levels. The study demonstrated the interest of NMF for urban sound environments and com-

pared the benefits of each approach on a validation corpus, designed to . However, it has now to face to real urban sound scenes.

ML : la fin de ce par est a revoir car c'est le point crucial du papier. Bien expliquer le corpus du papier d'avant, son usage, les questions qui restent en suspens et en quoi on passe ici a l'Altape superieure en y repondant.

In this paper, the NMF framework is applied on a corpus of simulated sound scenes, generated based on annotated urban recordings, and whose degree of realism is validated through a perceptual test. The NMF framework and its different optimization strategies 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 [28] and popularized by Lee and Seung [21]. 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 the dimensions  $F \times K$  and  $K \times N$  respectively.

$$\mathbf{V} \approx \mathbf{WH}. \quad (1)$$

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 Frequency 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.

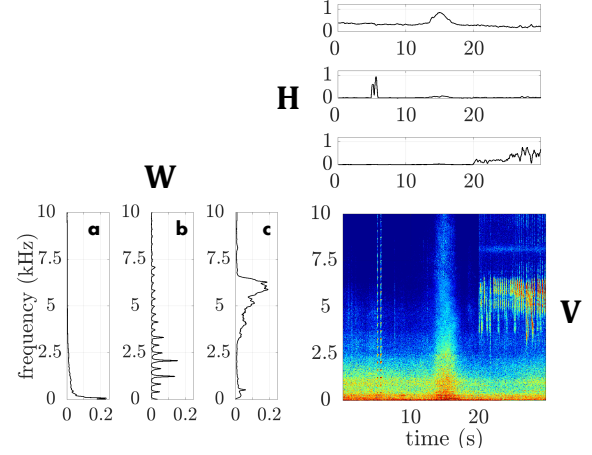


Figure 2: NMF decomposition of an audio spectrogram composed of 3 elements ( $K = 3$ ): passing car (a), car horn (b) and whistling bird (c).

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

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

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

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

$d_{\beta}(x|y)$  is usually chosen as a  $\beta$ -divergence [29], a sub-classes belonging to the Bregman divergences [30] which include 3 specific divergence calculations: the Euclidean distance (eq. 4a), the Kullback-Leibler divergence (eq. 4b) and the Itakura-Saito divergence (eq. 4c):

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

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 [31] or Projected Gradient [32] have been considered. The most commonly used algorithm is the Multiplicative Update [33]. The latter method is chosen here, as it ensures non-negative results and the convergence of the results [29].

### 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) [29].

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

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  $\beta > 2$ . The product  $A \otimes B$  and  $A/B$  symbolized 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 4.3. 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 it which is not possible in an urban environment. However, in our application setting, the main source of interest, *i.e.* the traffic, can be known *a priori*. **ML : manque une phrase** semi-supervised NMF (SEM-NMF) [34] is considered to better take into account the interfering sound sources. To do so,  $\mathbf{W}_{F \times (K+J)}$  is decomposed into two distinctive matrices:  $\mathbf{W} = [\mathbf{W}_s \mathbf{W}_r]$  where  $\mathbf{W}_{sF \times K}$  is a fixed part of  $\mathbf{W}$  composed of audio spectra and  $\mathbf{W}_{rF \times J}$ , a mobile part which is updated, see eq. 8a. Thus it is possible to include elements not present in  $\mathbf{W}_s$ . The dimension of  $\mathbf{W}_r$  is set up as  $J \ll K$  in order to best considered the

sound source present in  $\mathbf{W}_s$ .  $\mathbf{H}$  is then also decomposed in two matrices,  $\mathbf{H}_{(K+J) \times N} = \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_r \end{bmatrix}$ . The 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 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 to include in  $\mathbf{W}_r$  all the other sources 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 the mobile part  $\mathbf{W}_r$  gives more flexibility. 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 [35, 36] and speech content [25, 37].

### 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  $\mathbf{H}$ ,

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

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 closed to the traffic component. A cosine similarity  $D_\theta(\mathbf{W}_0\|\mathbf{W}')$  is computed for each element  $k$  as it is an invariant scale and a bounded method,

$$D_\theta(\mathbf{w}_0\|\mathbf{w}') = \frac{\mathbf{w}_0 \cdot \mathbf{w}'}{\|\mathbf{w}_0\| \cdot \|\mathbf{w}'\|} \quad (11)$$

where  $\mathbf{w}$  is a  $k$  element of  $\mathbf{W}$  of  $F \times 1$  dimension. When  $D_\theta(\mathbf{w}_0\|\mathbf{w}')=1$ , the element  $k$  from  $\mathbf{W}'$  is equal to the in  $\mathbf{W}_0$ . If  $D_\theta(\mathbf{w}_0\|\mathbf{w}')=0$ , the element is fully different. Next, the similarities are sorted in descending order. The extraction of traffic elements in  $\mathbf{W}'$  is carried out by a hard thresholding method [38]. It consists in weighting in a binary way the traffic elements  $\mathbf{W}'$  according to  $D_\theta(\mathbf{w}_0\|\mathbf{w}')$  and a threshold value such as:

ML : on pourrait tout a fait imaginer l'ajout d'un dictionnaire random pour se rapprocher de la semi.

$$\mathbf{w}_{traffic} = \alpha_k \mathbf{w}' \quad (12)$$

with

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

ML : faire un par resumanant les prop des trois methodes avec un accent sur l'utilisation de la connaissance a priori.

These methods are applied on simulated sound scenes in order to compare the estimated sound levels with a

non ambiguous reference. In [27], the sound corpus was composed of a mix of traffic samples with specific sound classes (*alert, animals, climate, human, mechanics, transportation*). Here, in order to implement this method in embedded sensors, a new and more realistic sound corpus by considering more realistic urban sound environments with many kind of sound sources.

### 3. Design of realistic urban sound scenes

ML : Il faut faire une figure et un paragraphe resumanant la figure. Voici un essai (je pense d'ailleurs que cela devrait faire partie des claims dans l'intro) : In order to validate the above described methods, the evaluation corpus is both required to be realistic and non ambiguous in terms of traffic sound level. The former calls for the use recordings of urban sound environments, and the latter imposes the use of controled use of sound sources with known properties: onset and offset time location, and volume. Indeed, the precise annotation of the sound level of a given source of interest in a polyphonic mixture is not feasible. We thus propose to design such evaluation corpus based on two corpora. The first, called the reference corpus, is a set of recordings of urban sound environments, from which a sort of "score" is extracted manually, *i.e.* which sound source is present from this time to this time, at which sound level. The second, called the elemental corpus, is a set of monophonic recordings of sound sources that are occurs in the scenes of the reference. The recordings of the elemental corpus are not taken from the reference corpus. The evaluation corpus is then built by sequencing recordings of the elemental corpus according to the scores of the reference corpus. (Le reste est a reecire avec les termes proposÃs)

More precisely, the urban sound scenes are taken from 76 recordings from 2 to 5 min, achieved in the 13th district of Paris (France) at 19 different locations <sup>1</sup>, which cover

<sup>1</sup>Recordings were made as part of the Grafic project funded by Ademe



Figure 3: Walked path with the 19 stop points [39].

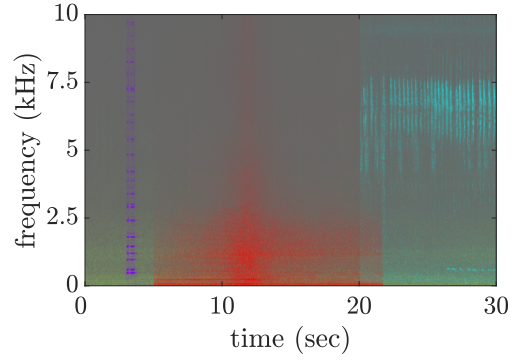


Figure 4: Spectrogram of a simple scene created with the *SimScene* software with a sound background (road traffic in green) and 3 sound events (car horn in purple, passing car in red and whistling bird in blue).

four various sound environments (Figure 3), see Figure 3.

A complete description of the experimental protocol can

be found in [39]. Two of the 76 recordings are rejected

for the analysis because the audio files were corrupted, re-

sulting in 74 valid audio files assumed as representative

of the variety of sound environments. The recordings are

listened and categorized within four different sound envi-

ronments, as proposed in [40]: park (P, 8 audio files with a

cumulative duration of 16min01), quiet street (Q, 33 audio

files with a cumulative duration of 77min27), noisy street

(N, 24 audio files with a cumulative duration of 56min10)

and very noisy street (vN, 9 audio files with a cumulative

duration of 21min42). Then, each audio file is annotated,

annotating the start, end time and level of each

sound event along with its sound class.

ML : cette phrase

ne dis probablement pas ce que tu veux dire annotation et

Transcription sont des termes equivalents, The aim of the

annotation phase is next to transcribe the recordings, in

order to obtain simulated sound scenes with the same ML

: ca peut preter a confusion en implicitant une distribution

statistique : distribution of sound events as the recordings

and therefore as close as possible to the realistic scenes.

### 3.1. Generation of the evaluation corpus

The sound scenes are generated with the *SimScene* soft-  
ware<sup>2</sup>, [41], which is a simulation software generating

<sup>2</sup>Open-source project available at: <https://bitbucket.org/mlagrange/simscene>

monaural sound mixtures in wav format with a 44.1 kHz  
sampling rate from an isolated sound database. This soft-  
ware have already been used in a wide range of experi-  
ments for sound detection algorithm assessment [42, 43].

ML : Pas clair The control of high level parameters can  
be handled by the user as the sound class occurrence, the  
time between each sample of one sound class, the ratio  
between the sound level of an event class with the back-  
ground (i.e the *event background ratio* shorten *ebr*). . . Each  
parameter is completed with a standard deviation to bring  
random behavior. It allows too the design of a sound mix-  
ture from an annotation text file. A remplaceur par ceci:  
The *SimScene* software allows for several type of sequenc-  
ing, from abstract ones (time indexes and amplitudes are  
drawn from random distributions) to precise ones, where  
the time indexes and amplitudes for each event is set by  
the user. The latter type is considered in this study. As  
output, *SimScene* generates an audio of the global sound  
mixture and an audio for each sound class present in the  
scene, which makes it possible to know their exact contri-  
butions in the scene, see Figure 4.

To transcribe the recordings in simulated scenes, a high  
quality sound database (wav format, 44.1 kHz sampling  
rate, high *Signal Noise Ratio*) has been built-up from au-  
dio samples found online (*freesound.org*) or with the help



of an already existing sound database [44]. The sound database 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, foot-step, metallic noise, voice ...*) and the *background* (or *texture*) category gathering 154 long duration sounds ( $\approx 1m30$ ), 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 has been made on the Ifsttar's runway. The recordings has been made for 4 cars (Renault Scenic, Renault M3gane, Renault Clio and Dacia Sandero), at different speeds and gear ratios. Overall, 103 car passages have been recorded. **ML : ajouter une description sur les probl3matiques d'overfitting et de generalisation. Cela doit etre un rappel d'un texte ecrit dans la partie nmf. A ce titre, cela pourrait etre une extension de l'etude, on en reparle** The audio samples of the first two cars (Renault Scenic and Renault M3gane) are included in the *SimScene*'s sound database (50 audio files in total). The last 53 audio samples are dedicated to the dictionary design as part of NMF, see section 4.3. A full description of the recordings can be found in [45].

With this built-up database, the *SimScene* software and the annotations of the recordings, 74 simulated sound scenes are generated, which have the temporal structure of the reference recordings. The *ebr* parameter is adjusted manually on each sound scene to be faithful compared to the recorded scenes. To validate this adjustment, the mean *Traffic Interfering Ratio* is calculated and summarized on Table 1. It expresses, on all the scenes of an sound environ-

Table 1: *mTIR* for each sound environment. **ML : les captions doivent etre self-contained, pas d'acronymes non evident**

	<i>mTIR</i> (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 ( $\pm 9.78$ )

ment, 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.

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

As it is the sound environment where the traffic is less present, the *mTIR* is always negative for the *park*. The *interfering* sound class is therefore the main sound source. In the case of the *quiet street*, the *TIR* can be positive or negative depending on the traffic presence on the scene. For the 2 others sound environments, when the traffic become the main sound source, *mTIR* is always positive.

### 3.2. Perceptual test

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

The perceptual test is conducted with a panel of 50 subjects 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 transcribed and recorded scenes. The total number of sound scenes tested is set at 40. The first half includes 20 30-seconds audio files, 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 transcribed scenes corresponding to the 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) [46] that determines the listening order of each participant based on fixed parameters (number of listeners, total number of audio, number of audio assessed by each listeners). The listening order is then built in a such way than each listener listens the same amount of recorded scenes and transcribed scenes to avoid statistical bias. The experimental design and the listening order per participant are performed with the package *sensMineR* on the *R* software [47].

The test was available online on the 8 February 2017 and the needed number of participant 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 <sup>3</sup>. 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. Figure 5 summarizes the score distributions of all the listeners for the recorded (rec.) and the transcribed (trans.) scenes.

The distributions of the notations according to the type are extremely similar. The mean score for the transcribed scenes is even superior to the recorded ones ( $m_{trans.} = 5.1$

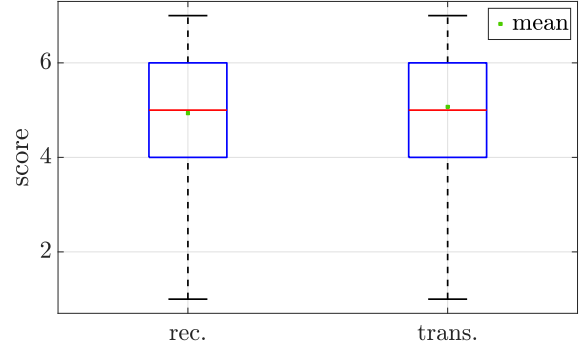


Figure 5: Box and whiskers plot of the rating of realism according to the type (recorded /transcribed). ML : Il faut tout expliciter, as tu vraiment besoin de toutes ces informations ?

( $\pm 1.6$ ),  $m_{rec} = 4.9$  ( $\pm 1.6$ )). These results confirm that the recorded and the transcribed scenes are perceived in a similar way by the panel. More details on the results can be found in [45].

ML : ce serait quand meme bien d'avoir un test stat, au moins un ttest

As the transcribed sound corpus is sufficiently close to the recordings, these sound mixtures can be used to assess the performances of NMF to estimate the traffic sound level.

## 4. Experimental protocol

The design of the experiment aims at evaluating in a meaningful manner the performance of the different systems. To do so, the 74 sound scenes, organized on 4 different types of sound environments are iteratively fed to the processing system 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. Metrics

The traffic sound levels,  $\tilde{L}_{p,traffic}$ , are compared to the exact values,  $L_{p,traffic}$ , through the Mean Absolute Error (MAE) [48]. The MAE expresses the quality of the

<sup>3</sup>The interface of the experiment is available <http://soundthings.org/research/xpRealism>



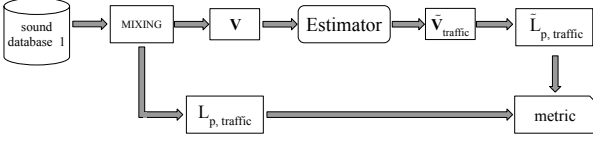


Figure 6: Bloc diagram of the experimental protocol.

long-term reconstruction of the signal and consists in the average of the absolute difference between the exact and the estimated sound levels,

$$MAE = \frac{\sum_{i=1}^M |L_{p,traffic}^i - \tilde{L}_{p,traffic}^i|}{M}. \quad (15)$$

It is then possible to express the  $MAE$  error for each unique setting but, also, to average this metric according to the 4 sound environments to be able to estimate the optimal setting that offers the lowest error for all the sound environments:

$$mMAE = \frac{\sum_{i=1}^4 MAE_i}{4}, \quad (16)$$

where the other experimental factors (method,  $f_c$ ,  $\mathbf{K}$ ,  $\mathbf{w}_t$ ,  $\beta$ , threshold value  $\mathbf{t}_h$ ) are fixed.

#### 4.2. Methods

As the road traffic is mainly composed of a low frequency content, a frequency low-pass filter (LP filter) is considered as baseline. It consists in filtering the sound scenes at different cut-off frequencies  $f_c \in \{500, 1, 2k, 5k, 10k, 20k\}$  Hz. The remaining energy located in the pass-band is assimilated to road traffic,

$$\tilde{\mathbf{V}}_{traffic} = \mathbf{V}_{f_c}. \quad (17)$$

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

#### 4.3. Dictionary building for NMF methods

The dictionary is designed from a second sound database specially dedicated to this task to prevent any overfitting

issues. It contains 53 audio files of the 2 cars (Dacia Sandero and Renault Clio) recorded on the runway, see part 3.1. Those recordings have not been considered for the creation of the evaluation dataset.

First the spectrogram of each audio file is computed (window  $w = 2^{12}$  with 50 % overlap). The spectrogram is then cut in multiple temporal frames of  $w_t = \{0.5, 1\}$  second duration. In each cut spectrogram, 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 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 for  $\mathbf{w}_t \in \{0.5, 1\}$  second, 2218 and 1109 elements.

A  $\mathbf{K}$ -means clustering algorithm is applied 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  $\mathbf{K}$  centroids are taken as the elements of  $\mathbf{W}$ . **ML : Vraiment utile ? Furthermore, the case where each audio generates one spectrum from its spectrogram is added ( $\mathbf{w}_t = all$ ). In this case, the dictionary elements are based on the spectral envelope of the audio samples. With only 53 audio samples, the number of elements in  $\mathbf{W}$  with the  $\mathbf{K}$ -mean clustering algorithm is reduced to  $\mathbf{K} \in \{25, 50\}$ .**

The obtained dictionary is expressed in third octave bands to decrease the dimensionality. **ML : Citer ou enlever, A previous experimental validation revealed that the use of third octave bands does not impact the performances of the chosen estimator in this study.** Each basis of  $\mathbf{W}$  is then normalized as  $\|\mathbf{w}_k\| = 1$  where  $\|\bullet\|$  is the  $\ell_1$  norm. Table 3 summarizes the different modalities of the two experimental factors ( $\mathbf{K}$  and  $\mathbf{w}_t$ ).

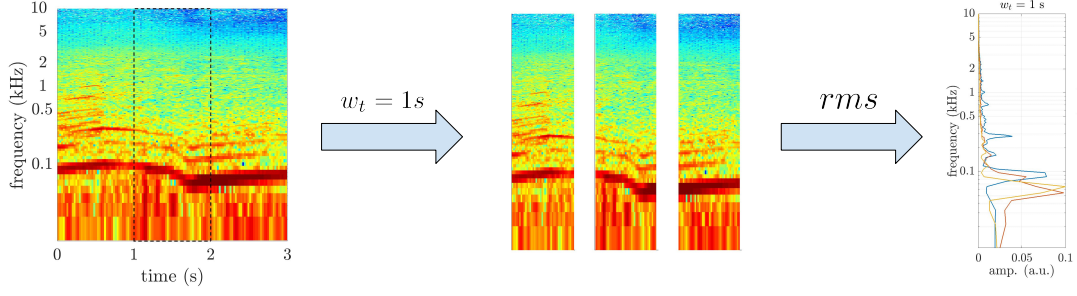


Figure 7: Dictionary building on a 3 second example of a passing car with  $w_t = 1$  second. In this case, the dictionary  $\mathbf{W}$  is made of 3 spectra, each representative of a texture frame of  $w_t$  duration.

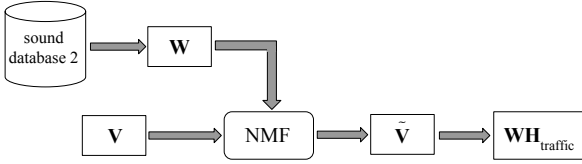


Figure 8: Bloc diagram of the NMF estimator.

#### 4.4. NMF experimental factors

NMF is performed for 3  $\beta$ -divergences:  $\beta = 2$  (euclidean distance),  $\beta = 1$  (Kullback-Leibler divergence) and  $\beta = 0$  (Itakura-Saito divergence). The spectrogram  $\mathbf{V}$  and the dictionary  $\mathbf{W}$  are represented in a logarithmic scale

through a third band octave 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 experimental factor forms an experimental setting. For the filter estimator, 24 settings are computed ( $4 \times 6$ ). For SUP and SEM-NMF, 240 settings are computed ( $4 \times 2 \times (2 \times 4 + 1 \times 2) \times 3$ ). Finally, for TI-NMF, the number of settings is much higher (3720) due to the higher cardinality of the set of threshold values ( $4 \times 1 \times (2 \times 4 + 1 \times 2) \times 3 \times 31$ ). The experimental factors and their different

modalities are displayed on Tables 2 and 3.

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,

$$\tilde{L}_{p,traffic} = 20 \log_{10} \frac{p_{rms}}{p_0}, \quad (18)$$

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.

## 5. Results and discussion

Table 4 summarizes the lowest  $mMAE$  errors according to the method (LP filter, SUP-NMF, SEM-NMF and TI-NMF) and  $\beta$  with the others corresponding modalities.

The LP filter with  $f_c = 20$  kHz cut-off frequency is equivalent to considerate the sound level of the entire scene without specific distinction between the sound sources. The error is then important with a high standard deviation ( $mMAE = 3.76 (\pm 4.35)$  dB). The lowest error for a LP filter is obtained with  $f_c = 500$  Hz ( $mMAE = 2.14 (\pm 1.83)$  dB). **ML : It allows a balance between the discarded and remaining energy through the sound environments.**

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  $\beta$  values. By adding the mobile part  $\mathbf{W}_r$  in the dictionary, SEM-NMF with  $\beta = 0$  and  $\beta = 1$  allows a

Table 2: Experimental factors and their modalities for the frequency low-pass filter baseline estimator.

experimental factors		modalities					number of modalities
sound environment	park	quiet street	noisy street	very noisy street	4		
	$f_c$ (kHz)	0.5	1	2	5	10	20

Table 3: Experimental factors and their modalities for the NMF estimator.

experimental factors	modalities				number of modalities
sound environment	park	quiet street	noisy street	very noisy street	4
method	SUP NMF	SEM NMF	TI NMF		3
$w_t$ (s)	0.5	1	<i>all</i>		3
$K$	25	50	100	200	4
$\beta$	0	1	2		3
hard threshold $t_h$	from 0.30 to 0.60 with a 0.01 step				31

lower error than 500 Hz LP filter with a reduced standard deviation especially for  $\beta = 1$  ( $mMAE = 1.94 (\pm 0.38)$  dB).

TI-NMF is the approach with the lowest global error ( $< 1.50$  dB). The best result is obtained for TI-NMF ( $MAE = 1.24 (\pm 1.24)$  dB) with  $\beta = 2$ ,  $K = 200$ ,  $w_t = 0.5$  s and as threshold value  $t_h = 0.32$ . This combination of settings offers the most fitted method to be adapted to all the sound environments. Furthermore, on the dictionary creation, except SEM-NMF where all the best methods according to  $\beta$  use the same dictionary, no specific dictionary form through all the method, is used. SUP and SEM-NMF privilege a high number of element ( $K = 200$ ) with a fine description of the audio samples ( $w_t \in \{0.5, 1\}$  second). For TI-NMF, the composition of the initial dictionary is more diverse both on the  $K$  value and on the finesse of the description ( $w_t \in \{all, 0.5\}$  second).

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 the same error evolution: a decrease of the error with the increase of the traffic predominance. On contrary, SEM-NMF show 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 remaining energy as the traffic component, no distinction is made between the different sound sources not related to the traffic component. In the opposite, for noisy and very noisy environments, the performances of the LP filter are good ( $MAE < 1$  dB). The errors are then due to a high deletion of the traffic energy by the filter while it becomes the main sound source.

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.

Table 4: Best  $mMAE$  errors according to the experimental factors  $\beta$  and  $method$  (in bold letter, the lowest error).

method	$f_c$ (kHz)	$\beta$	$\mathbf{K}$	$\mathbf{w}_t$ (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)
<b>TI-NMF</b>	-	<b>2</b>	<b>200</b>	<b>0.5</b>	<b>0.32</b>	<b>1.24 (<math>\pm</math> 1.24)</b>

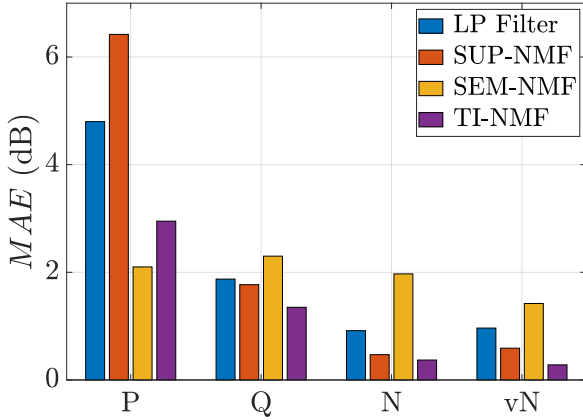


Figure 9:  $MAE$  errors according to each sound environment for the best combination of the LP filter ( $f_c = 500$  Hz), SUP-NMF ( $\beta = 2$ ,  $\mathbf{K} = 25$ ,  $\mathbf{w}_t = 1$  second), SEM-NMF ( $\beta = 1$ ,  $\mathbf{K} = 200$ ,  $\mathbf{w}_t = 1$  second) and TI-NMF ( $\beta = 2$ , 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

With this method, as NMF minimizes the cost function, eq. 2, the dictionary's elements are used to model the other sound sources. In 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 advanta-

geous 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 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.

ML : Pas clair, enlever ? : Besides having the lowest  $mMAE$  error, TI-NMF presents the most performing results. The park environment is the only case where an other NMF approach out-performed TI-NMF ( $MAE = 2.95$  dB). In this sound environment, the traffic dictionary is then composed, in average, of more than half the 200 elements in  $\mathbf{W}'$  (136 elements). For the rest of the sound environments, TI-NMF has the lowest error. For very noisy environment the error is even very low ( $MAE = 0.28$  dB). In this case, in average,  $\mathbf{W}'$  is composed of 198 traffic elements. This method out-performed SUP-NMF since, as  $\mathbf{W}_0$  is updated, the final dictionary  $\mathbf{W}'$  is then most suited to the sound scene than a fixed dictionary.

Considering a unique dictionary fitted to the sound scene under evaluation thus makes TI-NMF very effective

when traffic is predominant while, by the use of the thresholding, it discards the elements of the dictionary that deviated too much when the traffic is less present.

## 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 method of choice. If the sound environment can be identified through a prior analysis, or based on positioning data [49, 50], 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 would prove useful in the context of multi-source noise mapping that is gaining interest [51, 52]. Finally, the selected parameters stand for the corpus of sound mixtures of this study. 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 <sup>4</sup>.

## References

- [1] C. Nugent, N. Blanes, J. Fons, et al., Noise in europe 2014, European Environment Agency 10 (2014) 2014.
- [2] S. Kephelopoulou, M. Paviotti, F. A. Ledee, Common noise assessment methods in europe (cnossos-eu) (2012).
- [3] H. Van Leeuwen, S. Van Banda, Noise mapping-state of the art-is it just as simple as it looks?, Proceedings of EuroNoise 2015.
- [4] 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.
- [5] 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.
- [6] 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 through AIP Publishing LLC, Acoustical Society of America and the European Acoustics Association, Boston, United States, 2017, p. 3808.

<sup>4</sup><https://github.com/jean-remyGloaguen/articleApplied2018>

- [7] 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.
- [8] 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.
- [9] P. Maijala, Z. Shuyang, T. Heittola, T. Virtanen, Environmental noise monitoring using source classification in sensors, *Applied Acoustics* 129 (2018) 258–267.
- [10] X. Sevilano, J. C. Socoró, F. Alías, et al., DYNAMAP—development of low cost sensors networks for real time noise mapping, *Noise Mapping* 3 (1).
- [11] 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.
- [12] N. Lefebvre, X. Chen, P. Beuseroy, M. Zhu, Traffic flow estimation using acoustic signal, *Engineering Applications of Artificial Intelligence* 64 (2017) 164–171.
- [13] P. Mioduszeński, J. A. Ejsmont, J. Grabowski, D. Karpiński, Noise map validation by continuous noise monitoring, *Applied Acoustics* 72 (8) (2011) 582–589.
- [14] 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.
- [15] R. Ventura, V. Mallet, V. Issarny, et al., Estimation of urban 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.
- [16] 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.
- [17] B. Defreville, F. Pachet, C. Rosin, P. Roy, Automatic Recognition of Urban Sound Sources, *Audio Engineering Society*, 2006.
- [18] 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.
- [19] M. Cowling, R. Sitte, Comparison of techniques for environmental sound recognition, *Pattern Recognition Letters* 24 (15) (2003) 2895–2907.
- [20] 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.
- [21] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* 401 (6755) (1999) 788–791.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] I. Satoshi, K. Hiroyuki, NMF-based environmental sound source separation using time-variant gain features, *Computers & Mathematics with Applications* 64 (5) (2012) 1333–1342.
- [27] 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, submitted for publication (2018).
- [28] 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.
- [29] C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the  $\beta$ -divergence, *Neural Computation* 23 (9) (2011) 2421–2456.
- [30] R. Hennequin, B. David, R. Badeau, Beta-Divergence as a Subclass of Bregman Divergence, *IEEE Signal Processing Letters* 18 (2) (2011) 83–86.
- [31] 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.
- [32] C. J. Lin, Projected Gradient Methods for Nonnegative Matrix Factorization, *Neural Computation* 19 (10) (2007) 2756–2779.
- [33] D. Lee, H. Seung, Algorithms for Non-negative Matrix Factor-



- ization, in: In NIPS, MIT Press, 2000, pp. 556–562.
- [34] H. Lee, J. Yoo, S. Choi, Semi-Supervised Nonnegative Matrix Factorization, *IEEE Signal Processing Letters* 17 (1) (2010) 4–7.
- [35] 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.
- [36] 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 Fundamentals of Electronics, Communications and Computer Sciences* E97.A (5) (2014) 1113–1118.
- [37] 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.
- [38] 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.
- [39] P. Aumond, A. Can, B. De Coensel, D. Botteldooren, C. Ribeiro, C. Lavandier, Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context, *Acta Acustica united with Acustica* 103 (1–1).
- [40] 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.
- [41] M. Rossignol, G. Lafay, M. Lagrange, N. Misdariis, SimScene: a web-based acoustic scenes simulator, in: *1st Web Audio Conference (WAC)*, 2015.
- [42] G. Lafay, M. Rossignol, N. Misdariis, M. Lagrange, J.-F. Petiot, 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.
- [43] 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.
- [44] 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.
- [45] J.-R. Gloaguen, A. Can, M. Lagrange, 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) (2017) 4044–4044.
- [46] J. John, T. J. Mitchell, Optimal incomplete block designs, *Journal of the Royal Statistical Society. Series B (Methodological)* (1977) 39–43.
- [47] S. Lê, F. Husson, SensoMineR: A package for sensory data analysis, *Journal of Sensory Studies* (2008) 14 – 25.
- [48] 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.
- [49] 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.
- [50] C. Lavandier, P. Aumond, S. Gomez, C. Domingues, Urban soundscape maps modelled with geo-referenced data, *Noise Mapping* 3 (1).
- [51] C. A. Aumond P., Jacquesson L., Probabilistic modeling framework for multisource sound mapping, submitted for publication, submitted for publication.
- [52] F. Aletta, J. Kang, Soundscape approach integrating noise mapping techniques: a case study in brighton, uk, *Noise Mapping* 2 (1).