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

Acoustic sensor networks are increasingly deployed in cities, and appear more and more as a possible tool to enrich modeled road traffic noise maps through data assimilation techniques. This, or more simply the validation of the modeled maps through measures, requires first being able to isolate from the measured sound mixtures the road traffic sound level. 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. A corpus of sound scenes is artificially built imitating real annotated recordings. The realism of the 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. Three variants of NMF are tested, namely supervised, semi-supervised, and threshold initialized NMF. While the semi-supervised approach is the most appropriated for park environments, threshold initialized NMF, which is developed especially for this research purpose, appears to be the best generic approach, allowing road traffic noise level estimation with mean 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

---

1. Introduction

Noise mapping has been recommended as a tool to tackle noise pollution, in response to the growing demand from urban dwellers for a better environment. 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 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 computations, 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 computer efficiency 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 a vector of uncertainty. Moreover,

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

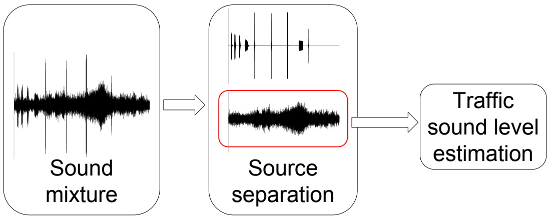

Figure 1: Block diagram of the blind source separation model

the produced aggregated indicators mask the sound levels evolution due to the traffic variations throughout the day.

Therefore, noise measurements are 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 protocol enables a priori combining measures and modeling 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. modify the acoustical parameters of the simulation thanks to noise measurements. Mallet et al. 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 [14] or the recognition [15] of sound events in environmental sound scenes. In these cases, a usually two-step schemes is followed were 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 . . .) [16, 17]. 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 [18], 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 sound events in order to discard them.

An other approach, followed in this paper, is to consider the Blind Source Separation paradigm, see Figure 1, which consists in the extraction of a specific signal inside a set of mixed signals. From the different existing methods (CASA, ICA), Non-negative Matrix Factorization (NMF) [19], seems the most relevant method for monophonic sensor networks. Dealing easily with the overlapping between the sound sources, this method approximes the magnitude spectrogram of a single signal by the product of two non negative matrices. A lot of applications can be found for musical [20, 21] and speech [22, 23] contents. Dedicated to sound separation, Immani and Kasaï used NMF in a two steps sound separation with the help of time variant gain features. A first study [] has been conducted, in which diverse NMF formula, 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 proved the interest of NMF for urban sound environments and compared the benefits of each approach. However, it has now to face to real urban sound scenes.

In this paper, the NMF framework is applied on a corpus of simulated sound scenes, generated based on annotated urban recordings, and whose realism is validated by a perceptual test. The NMF framework and its different implemented versions 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 through a perceptual test. Finally, in section 4 and 5, the experimental protocol and the results are exposed.

2. Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is a linear approximation method proposed by Paatero and Tapper [24] and popularized by Lee and Seung [19]. 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.

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

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 an audio file is considering, $\mathbf{V}$ 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 evolution of each spectrum, see Figure 2. Because of the non-negativity constraint, only additive combinations between the elements of $\mathbf{W}$ are considered. The dictionary is then composed of elementary elements providing a part based representation.

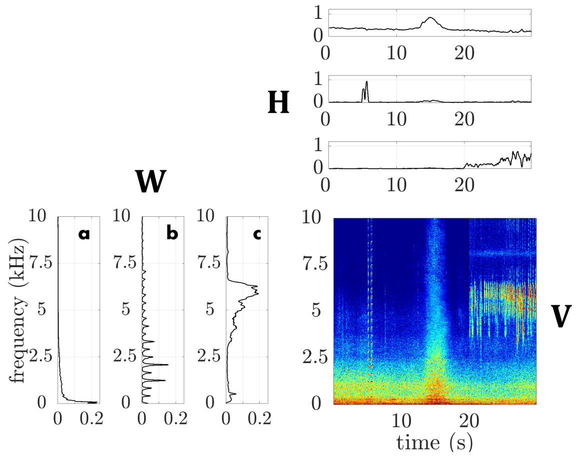

Figure 2: NMF for an audio sample with 3 elements ($K = 3$): passing car (a), car horn (b) and whistling bird (c)

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

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

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_β is mainly chosen as a β-divergence [25], a sub-classes belonging to the Bregman divergences [26] which include 3 specific divergence and distance 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 [27] or Projected Gradient [28] have been proposed. The most commonly algorithm used, and the chosen method here, is Multiplicative Update [29] as it ensures non-negative results and the convergence of the results [25].

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 determine and is updated at every iteration (eq. 5) [25].

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

with $\gamma(\beta) = \frac{1}{2-\beta}$, for $\beta < 1$, $\gamma(\beta) = 1$, for $\beta \in [1, 2]$ and $\gamma(\beta) = \frac{1}{\beta-1}$ for $\beta > 2$. The product $A \otimes B$ and A/B symbolized the Hadamard product and ratio.

Here, in an urban context, the sound sources are known and their audio samples can be obtained to learn $\mathbf{W}$, see section 4.1. 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:

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

2.2. Semi-supervised NMF

The main issue with the supervised approach is the limit imposed by $\mathbf{W}$: to be successful, all the acoustical sources must be considered in it which is not possible in an urban environment. As the main source that interests us can be known *a priori*, semi-supervised NMF (SEM-NMF) [30] is considered to better take into account the interfering sound sources. This approach proposes to decompose $\mathbf{W}_{F\times(K+J)}$ with 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 fixed audio spectra and $\mathbf{W_r}_{F\times J}$, 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 $J << K$ 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

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

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

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

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

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

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,

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

This approach as the advantage, with the add of the mobile part $\mathbf{W_r}$, to bring more flexibility and then to be more adaptive to the different urban sound environments. Applications of Sem-NMF can be found for musical [31, 32] and speech content [23, 33].

2.3. Thresholded Initialized NMF

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 closed 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 an invariant scale and a bounded method,

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

where $\mathbf{w}$ is a k element of $\mathbf{W}$ of $F \times 1$ dimension. When $D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right)$=1, the element k from $\mathbf{W'}$ is exactly similar than in $\mathbf{W_0}$. If $D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right)$=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 two thresholding methods. For both, the traffic elements are estimated by weighting $\mathbf{W'}$ according to $D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right)$ and a threshold value such as :

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

The first method is the hard thresholding [34]. It selects the traffic elements from $\mathbf{W'}$ in a binary way by comparing their similarities to a fixed threshold t_h

$$\alpha_k = \begin{cases} 1 & \text{iff} \quad D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right) > t_h, & \text{(13a)} \\ 0 & \text{else.} & \text{(13b)} \end{cases}$$

A second approach is based on firm thresholding [35] which considers two thresholds, $t_{f,1}$ and $t_{f,2}$ with $t_{f,1} > t_{f,2}$. The weight α_k is then defined as

$$\alpha_k = \begin{cases} 1 & \text{iff } D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right) > t_{f,1} & \text{(14a)} \\ 0 & \text{iff } D_\theta\left(\mathbf{w_0}\|\mathbf{w'}\right) \leq t_{f,2} & \text{(14b)} \\ \mathcal{N}\left(D_\theta(\mathbf{w_0}\|\mathbf{w'})\right) & \text{else.} & \text{(14c)} \end{cases}$$

where $\mathcal{N}\left(D_\theta(\mathbf{w_0}\|\mathbf{w'})\right)$ is a normalization between 0 and 1 of the similarity. The most similar elements to $\mathbf{W}_0$ are then keeping as they are whereas the elements whose the similarities are located between the two thresholds are then weighted. The allure of $D_\theta\left(\mathbf{W_0}\|\mathbf{W'}\right)$ and α are displayed in Figure 3.

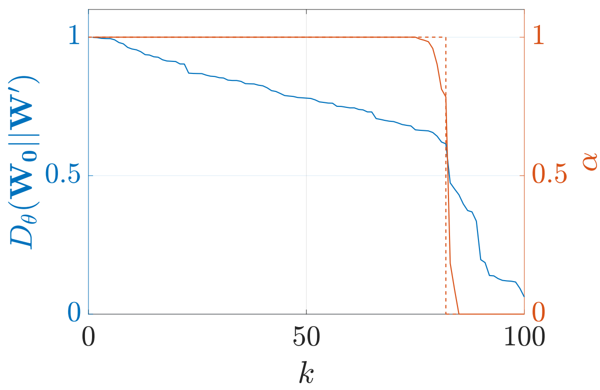

Figure 3: Evolution of the α weight according to $D_\theta\left(\mathbf{W_0}\|\mathbf{W'}\right)$ for hard thresholding with $t_h = 0.6$ (in dashed line) and firm thresholding with $t_{f,1} = 0.67$ and $t_{f,2} = 0.40$ (in full line)

These methods are applied on simulated sound scenes in order to compare the estimated sound levels with the exact solutions. In [], 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 more realistic sound corpus is generated based on urban recordings mixing all kind of sound sources.

3. Design of realistic urban sound scenes

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 [1], which cover four various sound environments (Figure 4). A complete description of the experimental protocol can be found in [36]. 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 sound environments. The recordings are listened and categorized

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

Figure 4: Walked path with the 19 stop points [36]

.

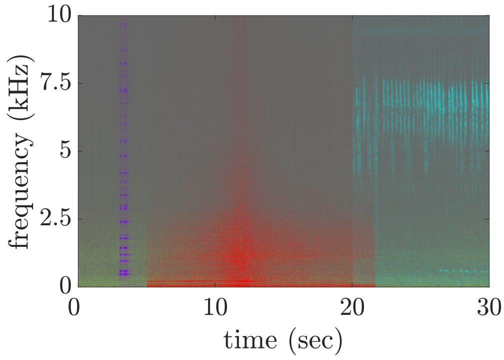

Figure 5: Spectrogram of a simple scene created with the *SimScene* software with a sound background (road traffic in green) and 3 sound events (car horn in purple, passing car in red and whistling bird in blue)

within four different sound environments, as proposed in [37]: park (P, 8 audio files), quiet street (Q, 35 audio files), noisy street (N, 23 audio files) and very noisy street (vN, 8 audio files). Then, each audio file is annotated, noting the start and end time of each sound event along with its sound class (described here below). The aim of the annotation phase is next to transcribe the recordings, in order to obtain simulated sound scenes with the same distribution of sound events as the recordings and therefore as close as possible to the realistic scenes.

3.1. Transcription of the recordings

The sound scenes are generated with the *SimScene* software[2], [38], which is a simulation software generating monaural sound mixtures in wav format with a 44.1 kHz sampling rate from an isolated sound database. This software have already been used in a wide range of experiments for sound detection algorithm assessment [39, 40]. The control of high level parameters can be handle by the user as the sound class presence, the time between each sample of one sound class, the ratio between the sound level of an event class with the background (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 mixture from an annotation text file. 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 contributions in the scene, see Figure 5.

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 audio samples found online (*freesound.org*) or with the help of an already existing sound database [41]. 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, footstep, metallic noise, voice* ...) and the *background* (or *texture*) category gathering 154 long duration sounds (≈ 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 present in multiple samples (*carHorn01.wav, carHorn02.wav* ...) 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

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

Senic, Renault Megane and Renault Clio, Dacia Sandero), for different speeds and gear ratios. In all, 103 car passages have been recorded. The audio samples of the first two cars (Renault Senic and Renault Megane) are included in the *SimScene*'s sound database (50 audio files totally). The last 53 audio samples are dedicated to the dictionary design as part of NMF, see section 4.1. A full description of the recording can be found in [42].

With this built-up database, the *SimScene* software and the annotations of the recordings, 74 simulated sound scenes are generated, which have the same temporal structure than the recordings. The *ebr* parameter is adjusted manually on each sound scene to be faithful compared to the recorded scenes. To validate its realism, the corpus of transcribed urban sound scenes is submitted to a perceptual test.

3.2. Perceptual test

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 mix 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 same 30-second transcribed scenes. In order to limit the duration of the test and to preserve the concentration of the listeners, each of them listens a subset of 20 sound scenes, which mix recorded and transcribed samples. Furthermore, all the scenes are normalized to the same sound level, chosen at 65 dB, to prevent the listeners from changing the sound level of their speakers.

The experimental design is elaborated following a partially Balanced Incomplete Block Design (PBIBD) [43] 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 avoide statistical biais. The experimental design and the listening order per participant are performed with the package *sensoMineR* on the *R* software [44].

The test was administered online on the 8 February 2017 and the 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. 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 (± 12) years old. 62% of the participants declared having no experience in the listening of urban sound mixtures. Figure 6 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 ($m_{trans.} = 5.1$ (± 1.6), $m_{rec} = 4.9$ (± 1.6)). These results confirm that all the recorded and the transcribed scenes are perceived in a

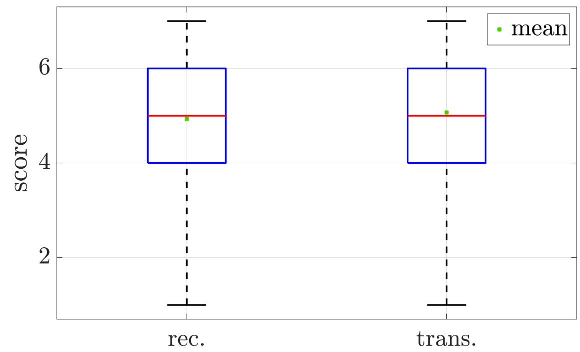

Figure 6: Box-and whiskers plot of the rating of realism according to the type (recorded /transcribed)

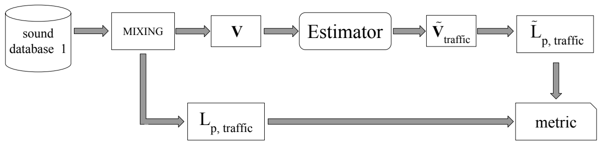

Figure 7: Bloc diagram of the experimental protocol

similar way by the panel. More details on the results can be found in [42].

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 experiment protocol consists in estimate, on the 74 scenes available classified on 4 sound environments, the equivalent sound level of the traffic of the entire scenes, $\tilde{L}_{p,traffic}$ (dB) and to compare them to their exact values given by the simulation process, $L_{p,traffic}$, see Figure 7.

As the road traffic is mainly composed of a low frequency content, a first estimator is considered through a frequency low-pass filter (LP filter). 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,

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

The second estimator is based on the three NMF formula presented in part 2 (see Figure 9). Between the dictionary building to the metric calculation, multiple experimental factors take part in these cases where each of them having different modalities.

4.1. Dictionary building

The dictionary building is designed from a second sound database specially dedicated to this task to prevent any overfitting issues. It contains the 53 audio files of the 2 other cars (Dacia Sandero and Renault Clio) recorded on the runway, see part 3.1.

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 8. From the 53 audio files, we obtain respectively for $w_t \in \{0.5,\ 1\}$ second, 2218 and 1109 elements. A K-mean clustering algorithm is applied to reduce these dimensions to $K \in \{25,\ 50,\ 100,\ 200\}$ in order to avoid redundant information and decrease the computation time. The obtained K clusters compose then the K elements of $\mathbf{W}$. Furthermore, the case where each audio generates one spectrum from its spectrogram is added ($w_t = all$). Here, by the added approach, the dictionary elements are based on the spectral envelops of the audio samples. For this case, having 53 audio samples, the number of elements in $\mathbf{W}$ with the K-mean clustering algorithm is reduced to $K \in \{25,\ 50\}$.

The obtained dictionary is expressed in third octave bands to decrease the dimensionality without deteriorate the spectral content. A previous experimental validation revealed that the use of third octave bands, instead of linear spectra, does not impact the performances of the chosen estimator in this study. Finally, each basis of $\mathbf{W}$ is normalized as $\|\mathbf{w}_k\| = 1$ where $\| \bullet \|$ is the ℓ_1 norm. Table 2 summarizes the different modalities of the two experimental factors (K and w_t).

4.2. 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

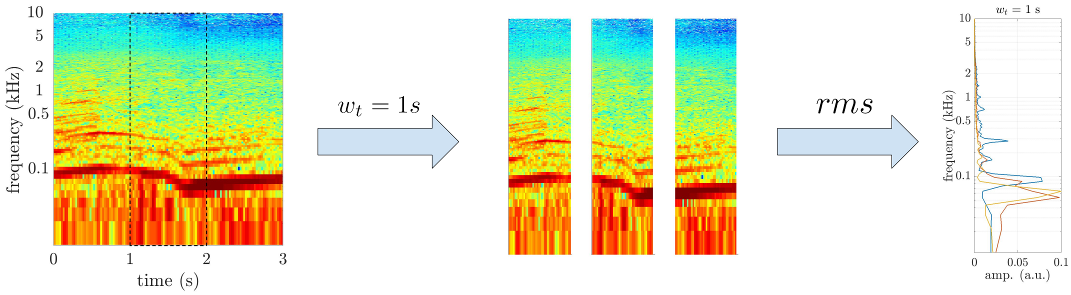

Figure 8: Dictionary building on a 3 second example of a passing car with $w_t = 1$ second. The dictionary are cut in 3 frames of w_t duration. On each the rms value is calculated to obtained 3 spectra destined to $\mathbf{W}$.

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

experimental factors	**modalities**						**number of modalities**
sound environment	park		quiet street	noisy street		very noisy street	4
$\mathbf{f_c}$ (kHz)	0.5	1	2	5	10	20	6

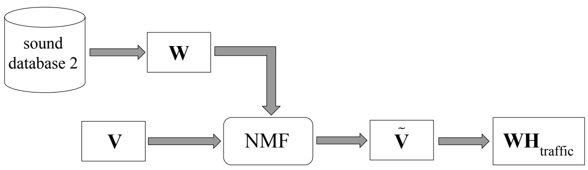

Figure 9: Bloc diagram of the experimental protocol

dictionary $\mathbf{W}$ are displayed 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 fixed at $J = 2$. For hard thresholding, the threshold value, t_h is defined between 0.30 and 0.60. For firm thresholding, the high threshold, $t_{f,1}$, is fixed between 0.45 and 0.60, the low threshold, $t_{f,2}$, is fixed between 0.10 and 0.45. All the intervals are defined with a 0.01 step. The cases $t_{f,1} < t_{f,2}$ are, naturally, discarded, see part 2.3. Each unique association of modalities between each experimental factor forms a setting. For the filter estimator, 24 settings are computed (4×6). 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 (115 320) due to the multiple possible combinations between the threshold values ($4 \times 1 \times (2 \times 4 + 1 \times 2) \times 3 \times (31 + (20 \times 31 + \sum_{i=15}^{30} i))$). The summarize of the experimental factors and their different modalities is displayed on Tables 1 and 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 deducted,

$$\tilde{L}_{p,traffic} = 20 \log_{10} \frac{p_{rms}}{p_0}, \tag{16}$$

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.

Table 2: Summary of the different experimental factors and their modalities taken into account in 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
$\mathbf{w_t}$ (s)	0.5		1	*all*	3
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
high firm threshold $\mathbf{t_{f,1}}$	from 0.30 to 0.60 with a 0.01 step				31
low firm threshold $\mathbf{t_{f,2}}$	from 0.10 to 0.45 with a 0.01 step				36

4.3. Metrics

The traffic sound levels, $\tilde{L}_{p,traffic}$, are compared to the exact values, $L_{p,traffic}$, through the Mean Absolute Error (MAE) [45]. The MAE expresses the quality of the long-term reconstruction of the signal and consists in the average of the absolute difference between the exact and the estimated sound levels,

$$MAE = \frac{\sum_{i=1}^{M} |L^i_{p,traffic} - \tilde{L}^i_{p,traffic}|}{M}. \tag{17}$$

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:

$$mMAE = \frac{\sum_{i=1}^{4} MAE_i}{4}, \tag{18}$$

where the other experimental factors (method, f_c, K, w_t, β, threshold value(s)) are fixed.

5. Results and discussion

Table 3 summarizes the lowest $mMAE$ errors according to the method (LP filter, SUP-NMF, SEM-NMF and Hard/Firm TI-NMF) and β 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 (± 4.35) dB). The lowest error for a LP filter is obtained with f_c = 500 Hz ($mMAE$ =2.14 (± 1.83) dB). 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 have 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 β = 0 and β = 1 allows a lower error than 500 Hz LP filter with a reduced standard deviation especially for β = 1 ($mMAE$ = 1.94 (± 0.38) dB).

Hard and Firm TI-NMF are the approaches with the

Table 3: Best $mMAE$ errors according to the experimental factors β and $method$ (in bold letter, the lowest error).

method	f_c (kHz)	β	**K**	$\mathbf{w_t}$ (s)	$\mathbf{t_h}$	$\mathbf{t_{f,1}}$	$\mathbf{t_{f,2}}$	$mMAE$ (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)
Hard TI-NMF	-	0	50	all	0.39	-	-	1.44 ($\pm$ 1.09)
Hard TI-NMF	-	1	200	0.5	0.35	-	-	1.40 ($\pm$ 1.22)
Hard TI-NMF	-	2	200	0.5	0.33	-	-	1.32 ($\pm$ 1.25)
Firm TI-NMF	-	0	50	all	-	0.45	0.32	1.42 ($\pm$ 1.11)
Firm TI-NMF	-	1	50	all	-	0.41	0.30	1.37 ($\pm$ 1.20)
Firm TI-NMF	**-**	**2**	**200**	**1**	**-**	**0.38**	**0.29**	**1.31 ($\pm$ 1.26)**

lowest global error (< 1.50 dB). The best result is obtained for firm TI-NMF (MAE = 1.31 ± 1.26 dB) with $\beta = 2$, $K = 200$, $w_t = 1$ s and as threshold values $t_{f,1} = 0.38$ and $t_{f,2} = 0.29$. This combination of settings offers the most fitted method to be adapted to all the sound environments. One observes that for firm TI-NMF and for each β, the thresholds values $t_{f,1}$ and $t_{f,2}$ frame the hard threshold value t_h.

The elements located between these threshold values, included in $\mathbf{W}_{traffic}$, despite a highest distance with the traffic spectra, their impacts can be reduced by the weighting. Firm TI-NMF allows then to take into account more elements as traffic component than hard thresholding and then decrease the $mMAE$ error.

Furthermore, on the dictionary creation, except SEM-NMF where all the best methods according to β 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 sound environments, see Figure 10.

Except 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. 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 can be made between the different sound sources not related to the sound sources of interest, here traffic. 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

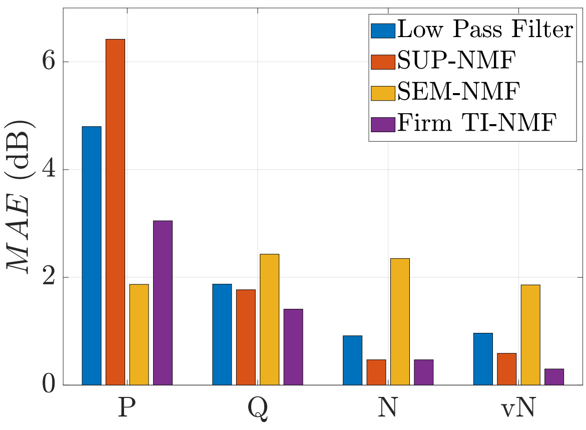

Figure 10: MAE errors according each sound environment for the best combination of the LP filter ($f_c = 500$ Hz), SUP-NMF ($\beta = 2$, $K = 25$, $w_t = 1$ second), SEM-NMF ($\beta = 1$, $K = 200$, $w_t = 1$ second) and firm TI-NMF ($\beta = 2$, $K = 200$, $w_t = 1$ second, $t_{f,1} = 0.38$ and $t_{f,2} = 0.29$)

traffic energy by the filter while it becomes the main sound source.

Despite a fixed dictionary composed of traffic spectra, SUP-NMF fails to identified 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. 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 advantageous for the *park* environment ($MAE = 1.87$ 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.

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 significantly firm TI-NMF ($MAE = 3.05$ 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, firm TI-NMF has the lowest error. For very noisy environment the error is even very low ($MAE = 0.30$ 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, $\mathbf{W}'$ is then most suited to the sound scene than a fixed dictionary. The advantage to have an unique dictionary suited to each sound scene makes TI-NMF very performing when traffic is predominant while, by the thresholding, it limits the dictionary derivation when the traffic is more quiet.

6. Conclusion

A non-negative matrix factorization frame has been applied as a source separation tool to estimate the traffic sound level from a corpus of urban sound scenes, which consists of scenes built artificially in which the level and position of each source is controled. In addition, the realism of the scenes has been verified thanks to a perceptual test. As a result, the proposed protocol mimics the outputs given by a deployed sensor network, but knowing the actual sound level of each source composing the sound mixture.

The results confirm the interest of this method on such sound environments. It easily takes into account the overlap between the multiple sound sources present in cities and is suited to monophonic sensor networks. Different NMF formula 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 failed on heavily traffic scenes. Finally, the proposed approach, namely Thresholded Initialized NMF, achieved the lowest error with an initialized updated dictionary. By considering the elements the most similar to traffic spectra with a firm thresholding, it makes it possible to adapt specifically the dictionary to the sound scenes. Consequently, in the case where the emplacement or the sound environment of the sensors cannot be identified (for instance within a mobile measurement framework), firm TI-NMF, with $\beta = 2$, $K = 200$, $w_t = 1$ second, $t_{f,1} = 0.38$ and $t_{f,2} = 0.29$, seems the best generic framework, for the sound corpus studied. But if the sound environment can be identified through a prior analysis, or based on positioning data [46, 47], it seems possible to adapt the proposed work by selecting the most efficient approach (for instance SEM-NMF within parks), 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, by replacing or adding elements in the built dictionary. This would prove useful in the context of muti-source noise mapping that is gaining interest [48, 49]. 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 test the robustness of the method, and select the most relevant approaches for specific sound environments (predominance of water or industrial sounds, rural environments . . .).

For reproducible purposes, the experimental protocol and the programs developed under the Matlab software are available online. This study proposes also a realistic urban sound corpus that could be greatly appreciate for research communities dedicated to the detection, identification or source separation tasks.

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) (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.

[7] N. Maisonneuve, M. Stevens, M. E. Niessen, L. Steels, Noisetube: Measuring and mapping noise pollution with mobile phones, in: Information technologies in environmental engineering, Springer, 2009, pp. 215–228.

[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. Sevillano, 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. Beauseroy, M. Zhu, Traffic flow estimation using acoustic signal, Engineering Applications of Artificial Intelligence 64 (2017) 164–171.

[13] P. Mioduszewski, J. A. Ejsmont, J. Grabowski, D. Karpiński, Noise map validation by continuous noise monitoring, Applied

Acoustics 72 (8) (2011) 582–589.

[14] 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.

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

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

[27] 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.

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

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

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

[31] 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.

[32] 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.

[33] 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.

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

[35] H. Gao, A. G. Bruce, Waveshrink with firm shrinkage, Statistica Sinica (1997) 855–874.

[36] 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).

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

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

[45] 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.

[46] 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.

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

[48] C. A. Aumond P., Jacquesson L., Probabilistic modeling framework for multisource sound mapping, submitted for publication, submitted for publication.

[49] F. Aletta, J. Kang, Soundscape approach integrating noise mapping techniques: a case study in brighton, uk, Noise Mapping 2 (1).