

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

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

Received: date / Accepted: date

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

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

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

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

1 Introduction

With the introduction of the European Directive 2002/EC/49, cities over 100 000 inhabitants have to produce road traffic noise maps. These maps depict the sound level distribution over the city and an estimation of the number of city dwellers exposed to high noise levels. These maps play both an important communication role and help drawing up action plans to reduce noise exposure. Road traffic noise maps are the result of a simulation process based on the estimation of the traffic density on the main roads and the use of sound propagation modelling. They express as output L_{DEN} and L_N values, which are *Day-Evening-Night* and *Night* equivalent A-weighted sound levels respectively. Although very useful, the produced noise maps introduce lot of uncertainty generated by the numerical tools [1] or by the different calculation methodologies used [2][3], despite the long data collection and calculation times. In addition, the usual road traffic noise maps are static, aggregating the exposure into indicators L_{DEN} and L_N , that ignore the sound levels evolution throughout the day. The use of acoustic measurements could facilitate their updating or even the generation of dynamic maps [4]. These measurements can be performed at fixed stations spread all over the cities [5] [6], which would make available of the long-term evolution of the traffic noise levels. It can also be performed with mobile stations [7] [8] covering a larger area with fewer sensors but also sparse time periods.

Currently, sensor networks in cities are spread for multiple applications (air quality assessment, measure-

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

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

ment of meteorological parameters ...), including the assessment of urban noise levels. The DYNAMAP project [9] studied the deployment and feasibility of such installations focusing on sensor installations on specific roads at the city scale in Milan and Rome [10]. The SONYC project (Sounds Of New-York City) aims to deploy a sensor network in New-York City for the purpose of monitoring constantly the noise pollution in the city [11]. In order to better know the urban sound environment, sensors are coupled with a detection tool that identifies the sound sources present [12]. In a similar way, but reduced to few neighborhoods with a denser network, the CENSE project¹ [13] aims to combine *in situ* observations, from a sensor network, and numerical data, from noise modeling, through data assimilation techniques.

Prior to data assimilation, the issue of the correct estimation of the traffic sound level from acoustic measurements is still unsolved [5]. Mainly because the urban sound environment is a complex environment gathering lots of different sounds (car passages, voices, whistling bird, car horn, airplanes...) that overlap. Consequently, the traffic sound level estimation based on measurements is not a trivial task. Many recent works have focused on the detection or recognition tasks of environmental sounds [14], [15], [16], [17]. A two-step process is generally followed : describe the audio files with a set of features (Spectral Centroid, harmonicity, Mel-Frequency Cepstral Coefficient ...) and classify them with the help of classifiers (Support Vector Machines, Gaussian Mixture Models, Hidden Markov Model, Artificial Neural Networks). A description of these features and classifiers can be found in [18] and their applications can be found in [19], [20], [21].

The main issue in the detection or recognition tasks is the overlap of environmental sounds. Although near major roads, traffic is predominant, there are many places where it overlaps with other sound sources which contribute significantly to the overall sound levels. To circumvent this issue, Socoró et al. propose to suppress time frames where there is significant overlap by considering an Anomalous Noise Events Detector [22]. It consists in detecting the unwanted sound sources from labeled recordings, *i.e.* that are not related to the traffic component. Those time frames are then discarded in order not to take them into account during the estimation of the traffic sound level. An alternative approach that we will follow in this paper is to consider the blind source separation paradigm to reliably estimate the traffic noise level, see Figure 1. It consists in separating the contribution of the traffic from the other sources within a polyphonic scene. One major advantage

of following such approach is that the estimate is continuously available, making the approach applicable in a wide range of urban areas, even where the traffic noise is relatively low compared to the remaining contributions.

In an urban environment context, source separation can be achieved with the help of acoustic microphone arrays and beamforming [23]. However, this approach requires spreading multiple microphones arrays in cities that is very expensive (even with low cost microphones) and time-consuming for calibration and maintenance. This method is then not considered here to be deployed all over cities. In the opposite, monophonic sensor networks need less microphones but the main challenge is to succeed to estimate correctly the road traffic from only one signal in which all kind of sound sources can be present. A convenient method for this is the Non-negative Matrix Factorization (NMF) technique [24]. When considering audio as input, it usually consists in approximating the magnitude spectrogram of an audio file by the product of two low rank matrices, one representing the components of interest and the other the contribution at a given time of those components to approximate the input magnitude spectrogram [25] [26] [27]. In the audio processing domain, NMF has already been employed for the source separation task of monaural signals of speech and music [28] [26]. By design, this method deals reasonably well with the overlapping sound sources as soon as the overlap can be resolved on the time/frequency plane. Closer to our application scenario, NMF has been considered by Innami and Kasai [29]. After having performed NMF on simulated audio files, they perform a source separation in two steps by 1) separating the sound background from the events and 2) by isolating the events using spectral features using a *k*-means procedure. We study in this paper different flavors of Non-Negative Matrix Factorization where the traffic component is considered in its entirety whether it is a sound background or an event. We demonstrate that supervised NMF and semi supervised NMF approaches have some interests but fail to give satisfactory results for the application at hand. We thus introduce another NMF scheme called thresholded initialized NMF that makes good use of prior knowledge about the source of interest, in our case the traffic noise, but also generalizes well to several kinds of urban areas and to traffic to interference ratio (*TIR*).

To perform the numerical experiments, we consider a corpus of simulated sound scenes created from a built-up sound database composed of a high number of diverse sound samples. The use of simulated sound scenes is mandatory for rigorous experimental validation as it

¹ <http://cense.ifsttar.fr/>

offers a high level of control on the design of the scenes and the knowledge of the exact contribution of the traffic component ($L_{p,traffic}$), which would be difficult to extract from a recording of an urban scene.

The remaining of the paper is organized as follows. Section 2 details the technical aspects of NMF and describes the 3 approaches considered in this paper to achieve the task at hand. Section 3 describes the corpus of environmental sound scenes and the experimental protocol setup. Section 4 presents and discusses the outcomes of the numerical results.

2 Non-negative Matrix Factorization

2.1 Description of NMF

Non-negative Matrix Factorization is a linear approximation method introduced by Lee and Seung, [24], which can be used to approximate the spectrogram $\tilde{\mathbf{V}}$ (obtained using a Short-Term Fourier Transform) of an audio file, \mathbf{V} , $\in \mathbb{R}_{F \times N}^+$ as :

$$\mathbf{V} \approx \tilde{\mathbf{V}} = \mathbf{W}\mathbf{H} \quad (1)$$

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

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

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

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

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

and usually belongs to the β -divergence class [30] in which the well known Euclidean distance (eq. 4a) and the Kullback-Leibler divergence (eq. 4b) belong

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

$$(4b)$$

To better take into account prior knowledge about the sources of interest, constraints (like the smoothness or the sparseness criteria [31]) can be added to the objective function.

Algorithms have been proposed to solve the minimization problem (2) iteratively such as the multiplicative update [32], the alternating least square method [33] or the projected gradient [34]. Here, the multiplicative update is chosen as it has been well studied in the literature and it ensures convergence of the results [35].

2.2 Supervised NMF

First, supervised NMF (Sup-NMF) is used: the *dictionary* includes audio spectra of urban sound sources. A lot of the different sound sources present in the urban environment are known. Their spectra can be obtained and be a basis of \mathbf{W} . The *activation* matrix is then the unknown variable to estimate. In the first iteration, \mathbf{H} is initialized randomly, then it is updated by the generic algorithm [35]

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

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 $A.B$ and A/B are respectively the Hadamard product and ratio. As in the supervised approach the indexes of traffic components in \mathbf{W} are known, the separation of the corresponding sound source is made by extracting the related basis and activators,

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

2.3 Semi-supervised NMF

The supervised approach is useful when prior knowledge can be assumed for all the sources in the mixture, which is not a reasonable assumption in our application scenario. To some extent, prior knowledge can be considered for the traffic but not for the numerous kind of interferences that can occur in a realistic scenario.

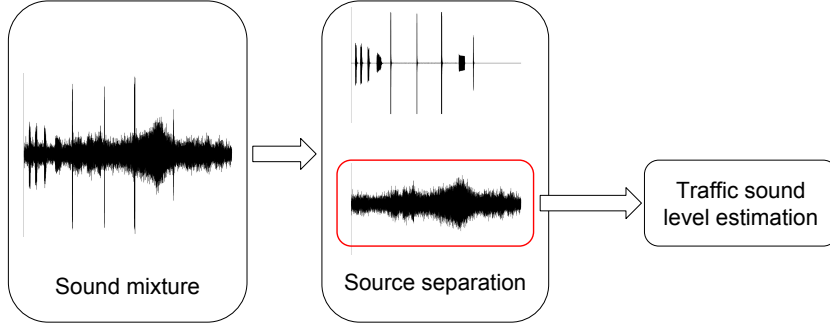


Fig. 1: Block diagram of the general source separation model.

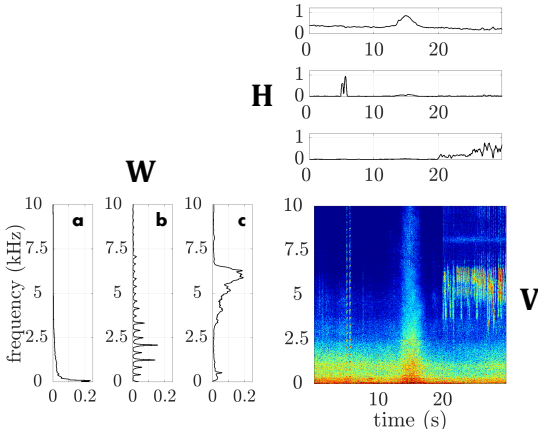


Fig. 2: Example of a simple NMF for urban sound mixture composed of 3 sound events (car passages, car horn and bird's whistles), \mathbf{W} is composed of 3 elements too ($K = 3$) which correspond in 3 audio spectra (car passages (a), car horn (b) and bird's whistles (c)).

To better take into account the diverse nature of urban scenes, semi-supervised NMF (Sem-NMF)[36] is a good candidate as it offers more flexibility. This method assumes a *dictionary* with a fixed part $\mathbf{W}_s \in \mathbb{R}_{F \times K}^+$, composed in our case of road traffic spectra, and with a mobile part, $\mathbf{W}_r \in \mathbb{R}_{F \times J}^+$ with $J \ll K$, that is updated during optimization. In the literature, J is set to a small number with respect to K so as to force the optimization to still consider the fixed part of the dictionary [37]. The aim is to include in \mathbf{W}_r the elements that are not related with the traffic. The problem (1) becomes

$$\mathbf{V} \approx \tilde{\mathbf{V}} = \mathbf{W}_s \mathbf{H}_s + \mathbf{W}_r \mathbf{H}_r \quad (7)$$

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

$$\mathbf{W}_r^{(i+1)} \leftarrow \mathbf{W}_r^{(i)} \cdot \left(\frac{\left[(\mathbf{W}_r \mathbf{H}_r^{(i)})^{(\beta-2)} \cdot \mathbf{V} \right] \mathbf{H}_r^T}{(\mathbf{W}_r \mathbf{H}_r^{(i)})^{(\beta-1)} \mathbf{H}_r^T} \right)^{\gamma(\beta)}, \quad (8a)$$

$$\mathbf{H}_r^{(i+1)} \leftarrow \mathbf{H}_r^{(i)} \cdot \left(\frac{\mathbf{W}_r^T \left[(\mathbf{W}_r \mathbf{H}_r^{(i)})^{(\beta-2)} \cdot \mathbf{V} \right]}{\mathbf{W}_r^T (\mathbf{W}_r \mathbf{H}_r^{(i)})^{(\beta-1)}} \right)^{\gamma(\beta)}, \quad (8b)$$

$$\mathbf{H}_s^{(i+1)} \leftarrow \mathbf{H}_s^{(i)} \cdot \left(\frac{\mathbf{W}_s^T \left[(\mathbf{W}_s \mathbf{H}_s^{(i)})^{(\beta-2)} \cdot \mathbf{V} \right]}{\mathbf{W}_s^T (\mathbf{W}_s \mathbf{H}_s^{(i)})^{(\beta-1)}} \right)^{\gamma(\beta)}. \quad (8c)$$

Applications of Sem-NMF for speech denoising from background noise or musical content can be found in [38] and [39].

2.4 Thresholded initialized NMF

As it will be demonstrated in the experimental results described in Section 4, those last approaches fail to provide consistent results in a wide range of urban areas and for different traffic preponderances due to generalization capabilities issues.

We therefore propose an alternative scheme based on the unsupervised NMF. Usually in unsupervised, \mathbf{W} , as \mathbf{H} , is initialized randomly. Here, as the concerned sound source is known and audio samples of car passages are available, an initial dictionary, \mathbf{W}_0 , is learnt by converting the audio files in the spectra domain; see part 3.2.1. Then NMF is performed where \mathbf{W} (eq. 9) and \mathbf{H} (eq. 5) are updated alternatively. \mathbf{W} is therefore updated by forcing its initiation with *a priori* knowledge but allowing it to adapt to the actual content of the scene under study,

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

After N iterations, a measure of similarity $D_\theta(\mathbf{W}_0||\mathbf{W})$ between \mathbf{W}_0 and the obtained dictionary \mathbf{W} for each element k is computed using a cosine similarity metric,

$$D_\theta(\mathbf{W}_{0k}||\mathbf{W}_k) = \frac{\mathbf{W}_{0k} \cdot \mathbf{W}_k}{\|\mathbf{W}_{0k}\| \cdot \|\mathbf{W}_k\|}. \quad (10)$$

$D_\theta(\mathbf{W}_{0k}||\mathbf{W}_k) = 1$ means that the k -th element of \mathbf{W} is identical to the k -th element of \mathbf{W}_0 . In the opposite, $D_\theta(\mathbf{W}_{0k}||\mathbf{W}_k) = 0$ means that the elements are completely different. This measure has the advantage to be bounded between 1 and 0 and to be invariant with respect to scale. The k values of $D_\theta(\mathbf{W}_0||\mathbf{W})$ are next sorted in descending order. The elements in \mathbf{W} that can belong to $\mathbf{W}_{traffic}$ are selected by a *hard thresholding* method. It is defined as:

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

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

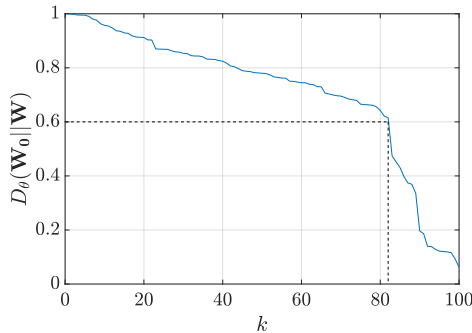


Fig. 3: $\mathbf{W}_{traffic}$ extraction from the sorted cosine similarity with a threshold $t = 0.6$. The 82-nd first elements are considered as traffic component.

This approach is named *Thresholded initialized NMF* (TI-NMF). Other thresholding methods as the *soft* [40] and the *firm* [41] have been investigated. A fast parametric study revealed that the *hard* thresholding method, as presented in Figure 3, was the most reliable approach.

3 Experimental protocol

In order to validate the usefulness of the proposed NMF scheme to estimate the road traffic noise levels, one need to compare the traffic noise level predicted by the algorithm to a reference level. The latter can hardly be

measured or even annotated from real life recordings. Thus, we propose to consider simulated sound scenes to assess the performance of the proposed NMF. This offers a controlled framework to design at low cost a wide diversity of sound environments in which all the traffic components are known, thus allowing the computation of the reference level.

3.1 Environmental sound scene corpus

The corpus is designed with the *SimScene* software². *SimScene* [42] is a simulator that creates monaural sound scenes in a .wav format by sequencing and summing audio samples that come from an isolated sound database. The simulator has been successfully considered for a wide range of experimental design for sound detection algorithm assessment [43] [44] [45].

This database is divided into two categories: *i*) the *event* category, which are the brief sounds (from 1 to 20 seconds) that are considered as salient including 245 sound event samples divided in 19 sound classes (*ringing bell, whistling bird, sweeping broom, car horn, passing car, hammer, drill, coughing, barking dog, rolling suitcase, closing door, plane, siren, footstep, storm, street noise, metallic noise, train, tramway, truck and voice*) and *ii*) the *background* or *texture* category that includes all the sounds that are of long duration and whose acoustic properties do not vary with respect to time. 154 sound samples that belong to this category are divided in 9 sound classes (*whistling bird, construction site noise, crowd noise, park, rain, children playing in schoolyard, constant traffic noise, ventilation, wind*). These sounds are in .wav format sampled at 44.1 kHz. The sound class *passing car* comes from 60 recordings of 2 cars (Renault Megane and Renault Scenic) made on the Ifstar's runway at different speeds with multiple gear ratio. The other audio files have been found online (*freesound.org*) and within the *UrbanSound8k* database [46]. Each sound class is composed of multiples samples (*bird01.wav, bird02.wav ...*) to allow some diversity in the resulting mixture, see Figure 4. The software allows the user to control some high level parameters (number of events of each class that appear in the mixture, elapsed time between each sample of a same class, presence of a fade in and a fade out ...) completed with a standard deviation that may bring some random behavior between the scenes. Furthermore, an audio file of each sound class present in the scene can be generated that allows us to know its exact contribution as well as a text file that summarizes the time presence of

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

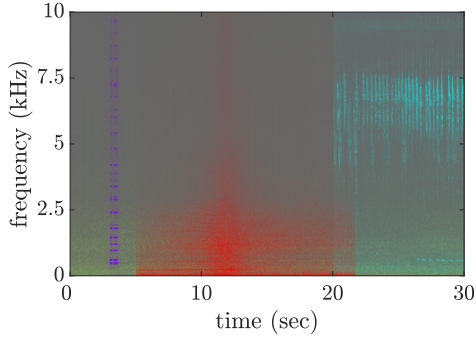


Fig. 4: Spectrogram of a sound scene created with *Sim-Scene* software with a sound background (road traffic in green) and 3 sound events (car horn in purple, car passage in red and whistling bird in cyan).

all the events.

This database allows the creation of a wide diversity of realistic urban sound scenes from the road traffic point of view [47]. A sound mixing corpus is composed of 6 sub-corpus of 25 audio files each lasting 30 seconds. Each sub-corpus is characterized by a specific generic sound class that summed with traffic will make the estimation of the traffic level more difficult. The classes are: *alert* (car horn, siren), *animals* (barking dog, whistling bird), *climate* (wind, rain), *humans* (crowd noise and voice), *mechanics* (metallic and construction site noises) and *transportation* (train, tramway and plane). In each file, traffic component is present as the sum of the background and event traffic sounds while the *interfering* sound class is the sound sources not related to it. To test different scenarios, each audio file is duplicated with the traffic sound level of the entire sound scene, $L_{p,traffic}$, fixed to a specific level according to the sound level of the interfering class, $L_{p,interfering}$, following the relation (12).

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

with the *Traffic Interference Ratio* $TIR \in \{12, -6, 0, 6, 12\}$ dB. When $TIR < 0$ dB, the traffic component is less present than the interfering class. In the opposite, for $TIR > 0$ dB, the traffic class is louder than the interfering class. The total number of scenes designed is 750 (6 sub-corpus \times 25 scenes \times 5 TIR values). The corpora are available for download at: <https://zenodo.org/record/1145855#.W12oPnkiGos>

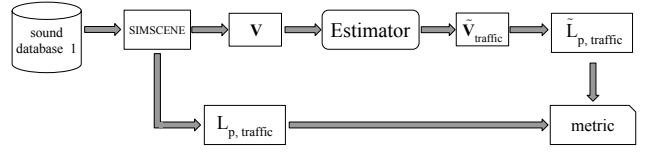


Fig. 5: Block diagram of the experience with the urban sound scenes sound mixture step and the estimation step. The estimator may be a frequency low-pass filter or NMF.

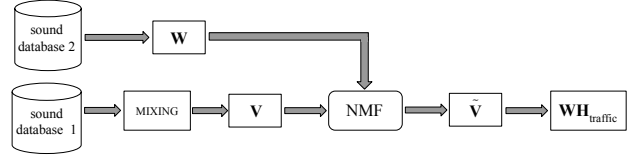


Fig. 6: Specific block diagram of the NMF estimator with the dictionary design composed from a second sound database.

3.2 Experiment

The experiment consists in estimating the road traffic sound level of the 6 environmental sound sub-corpus (*alert* (al), *animals* (an), *humans* (hu), *climate* (cl), *mechanics* (me), *transportation* (tr)) composed each of 25 audio files ($M = 25$) and for 5 TIR ($\{-12, -6, 0, 6, 12\}$ dB). The spectrogram \mathbf{V} of each sound scene is built with a window size $w = 2^{12}$ with a 50 % overlap, see Figure 5.

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

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

The second estimator is the proposed scheme, based on the three NMF approaches presented in Section 2. Multiples experimental factors are involved here between the dictionary learning and NMF (see Figure 6), each experimental factor having multiples modalities.

3.2.1 NMF Dictionary

In order to prevent potential overfitting issues, the dictionary is built from a separate sound database dedicated specifically to this task. It is composed of 53 audio files of passing cars. These recordings have been made

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
sub-classes	alert	animals	climate	humans	transportation	mechanics	6
TIR (dB)	-12	-6	0		6	12	5
f_c (kHz)	0.5	1	2	5	10	20	6

Table 2: Summary of the different experimental factors and their modalities taken into account in NMF estimator.

experimental factors	modalities						number of modalities
sub-classes	alert	animals	climate	humans	transportation	mechanics	6
TIR (dB)	-12	-6	0		6	12	5
method	Sup-NMF		Sem-NMF		TI-NMF		3
w _t (s)	0.5		1		all		3
K	25		50		100		4
β	1				2		2
threshold t	from 0.30 to 0.70 with 0.01 step						41

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

An example that illustrates the process can be found on Figure 7 on a 3 second extract of the spectrogram of a car passage, see Figure 7a. In the case where $w_t = 1$ second, 3 elements are therefore extracted from the

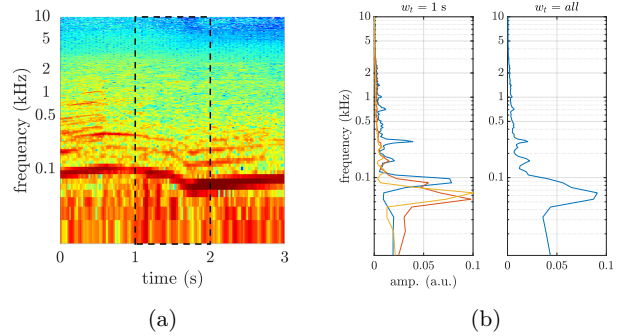


Fig. 7: Dictionary building on a 3 second extract of a car passage. In dashed lines, a 1 second w_t window (7a). With $w_t = 1$ second, 3 spectra are then generated and included in \mathbf{W} , while for $w_t = all$, the audio file is reduced into 1 spectral vector (7b).

spectrogram while in the case where $w_t = all$, all the spectrogram is reduced into one element, see Figure 7b.

The obtained dictionary is expressed with third octave bands and each basis vector of \mathbf{W} is normalized such as $\|\mathbf{W}_k\| = 1$ with $\|\bullet\|$ the ℓ_1 norm. Table 2 summarizes the experimental factors (K and w_t) for the dictionary building and their related modalities. This last step allows us to reduce the dimensionality while preserving a rich description of the spectral content. Experimental validation consistently showed that considering octave bands do not impact the performance of the estimator studied in this paper.

3.2.2 Experimental factors of NMF

Sup-NMF and Sem-NMF updates are computed for 400 iterations, which is sufficient to reach convergence. TI-NMF is performed on a lower number of iterations (60) to prevent \mathbf{W} to not deviate too much from the initial dictionary. The spectrogram \mathbf{V} , as the dictionary \mathbf{W} , is expressed with third octave bands ($F = 29$). This coarser method allows the reduction of the matrix size and decreases the computation time. But, must of all, by expressing the frequency axis on a log frequency axis, the low frequencies, where the traffic energy is focused, are described more finely than the high frequencies. For TI-NMF, the threshold t is set between 0.30 and 0.70 with a step of 0.01. Tables 1 and 2 summarize the experimental factors and their related modalities.

Considering the experimental settings derived from the different modalities of each experimental factor described in Table 1 between the 5 levels of *TIR*, the 6 sub-classes and the 6 cut-off frequencies f_c , 180 settings are performed ($6 \times 5 \times 6$). For Sup-NMF and Sem-NMF, according to Table 2, 1200 associations of factors are made where the 4 levels of K are associated with $w_t \in \{0.5 \text{ } 1\}$ second whereas only 2 levels of K (25 and 50) are associated with $w_t = \text{all}$, see part 3.2.1 ($6 \times 5 \times 2 \times (2 \times 4 + 1 \times 2) \times 2$). For TI-NMF, because of the high number of threshold t tested, 24600 combinations are computed ($6 \times 5 \times (2 \times 4 + 1 \times 2) \times 41$). In all, 25980 settings are performed.

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

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

where p_{rms} is the effective pressure deduced from the estimated traffic spectrogram $\tilde{\mathbf{V}}_{traffic}$ and p_0 is the reference sound pressure, $p_0 = 2 \times 10^{-5} \text{ Pa}$. For each setting of experimental factors, M values of $\tilde{L}_{p,traffic}$, corresponding to the M scenes, are then obtained and are compared to the M exact sound level, $L_{p,traffic}$.

3.2.3 Metrics

The performance of the road traffic sound level estimator is assessed through the calculation of one reference metric, the Mean Absolute Error (*MAE*) [48]. It expresses the quality of the long-term reconstruction of the signal and consists in the average over the M sound scenes of the absolute difference between the exact and estimated traffic sound level in dB,

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

4 Results

Table 3 summarizes, according to the 2 main factors (*method*, β), the lowest *MAE* error averaged on all sub-classes and all *TIR* (750 sound mixtures in all). For the low-pass frequency filters and each NMF approaches, the best parameter combinations are detailed according to the *TIR* in Table 4, and are expanded to the sub-classes in Figures 9a, 9b, 9c and 9d.

First, the errors produced by the filter are detailed. $f_c = 20 \text{ kHz}$ is equivalent to consider all the sound mixtures without distinction between traffic and others sound sources. Consequently, in low *TIR* (-12 dB and -6 dB), where traffic component is scarce, the error is more important than in high *TIR* (6 dB and 12 dB) where the traffic component is predominant. $f_c = 500 \text{ Hz}$ is the cut-off frequency with the lowest mean error obtained ($2.89 (\pm 2.84) \text{ dB}$). It is then considered as the baseline to compare the performances of NMF.

In low *TIR*, for *alert* and *animals*, which are sub-classes composed of higher frequencies, this filter is efficient as it removes these frequency components. For the other sub-classes where low frequency contents are present (storm for *climate*, voices in *humans*, planes, tramway and train in *transport* and ventilation noise in *mechanics*), the filter considers all the energy located in the pass-band and then does not dissociate the traffic element from the other sound sources. The errors are then nearly all superior to 4 dB and are overestimated, see Figure 8a. In opposite, in high *TIR*, the error is due to the energy removed from the traffic which has the consequence to underestimate the sound levels, see Figure 8b. The 500 Hz filter finds a balance between what is put aside in low *TIR* and what it is remained in high *TIR*.

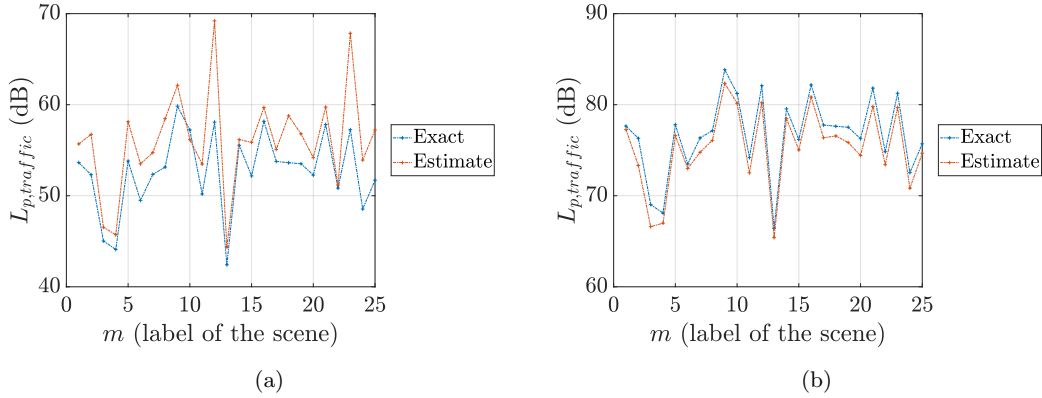
Compared with the filter errors, the choice of some NMF approaches makes it possible to decrease the error of the road traffic sound level estimation. The supervised approach is the only method that has an average error superior to the 500 Hz filter baseline. Sem-NMF and TI-NMF have better results. The lowest average error is obtained for TI-NMF, $2.19 (\pm 2.18) \text{ dB}$, for $\beta = 1$ and threshold $t = 0.42$ with the dictionary factors $K = 200$ and $w_t = 500 \text{ ms}$. On the other hand, the semi-supervised approach has a higher error but a lower standard deviation ($2.32 (\pm 1.26) \text{ dB}$).

Table 3: Best results for all the scenes according to the experimental factors β and *method* (in bold letter, the lowest error).

method	f_c (kHz)	β	K	w_t (s)	t	MAE (dB)
filter	20					4.69 (\pm 4.52)
filter	0.5					2.89 (\pm 2.84)
Sup-NMF		1	50	0.5		3.44 (\pm 3.70)
Sup-NMF		2	50	0.5		3.02 (\pm 3.33)
Sem-NMF		1	100	0.5		2.33 (\pm 1.10)
Sem-NMF		2	100	0.5		2.32 (\pm 1.26)
TI-NMF		1	200	0.5	0.42	2.19 (\pm 2.18)
TI-NMF		2	25	all	0.54	2.20 (\pm 2.26)

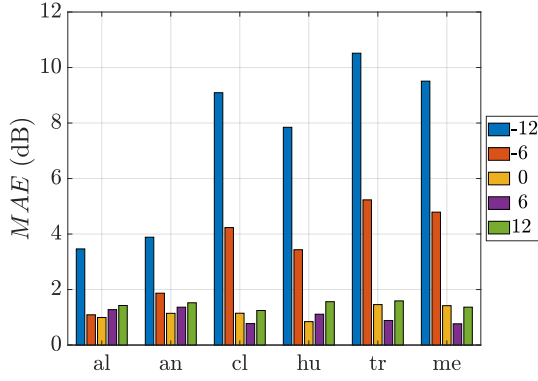
 Table 4: MAE error averaged on all sub-classes on each *TIR* for the best scenario according to each method.

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

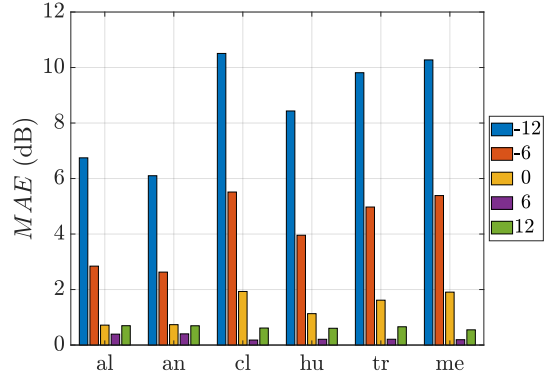

 Fig. 8: Estimated global sound levels of the traffic by the frequency low-pass filter with $f_c = 500$ Hz for the 25 scenes of the sub-classes *alert* at $TIR = -12$ dB (8a) and at $TIR = 12$ dB (8b).

According to Table 4 and Figures 9, the behavior between the 3 versions of NMF differs. In the case of Sup-NMF, it fails to improve the filtering performances despite good results in high *TIR*. The errors are too important for low *TIR* and this for all the sub-classes. This method reveals to be too rigid as \mathbf{W} is composed of fixed traffic spectra. In the case of low *TIR*, in the aim to reduce the objective function, see eq. 2, traffic elements are used whatever the sound event in the sound scene. Thus forcing the dictionary to be only traffic spectra is not a sufficient way to estimate correctly the traffic sound level, $\hat{L}_{p,traffic}$.

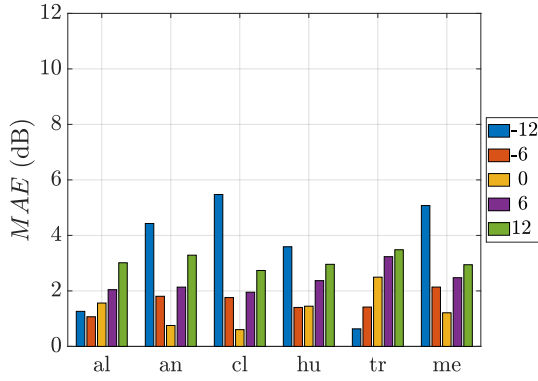
Consequently, with the addition of a mobile part in the dictionary, \mathbf{W}_r , the semi-supervised approach allows a better consideration of the interfering class in low *TIR*. It brings a significant decrease of the errors for low *TIR*. However, the relatively high degrees of freedom of Sem-NMF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



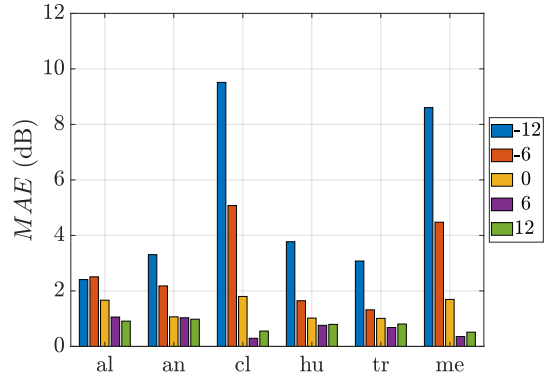
(a) MAE error for each TIR and sub-class with the frequency low-pass filter with $f_c = 500$ Hz.



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



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



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

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

Finally, TI-NMF with a threshold $t = 0.42$ and $\beta = 1$, offers the lowest average error (Table 3). Unlike Sup-NMF, where \mathbf{W} is fixed, and Sem-NMF, where only \mathbf{W}_r is updated, TI-NMF updates \mathbf{W} entirely to adjust prior knowledge to the scene under evaluation so as to adapt to the different sound environments, see Figure 10. The closest elements of the traffic component defined in \mathbf{W}_0 are then extracted to deduce the traffic signal. For low TIR , as the traffic sound class is not predominant, the final dictionary strongly differs from the \mathbf{W}_0 . With the thresholding, only a reduced number of basis vectors are considered as traffic components. In comparison to supervised results, this approach reduces significantly the error for the *human* and *transport* sub-classes. However, for *climate* and *mechanics*, the errors remain important as these interfering classes have similar spectral profiles when compared to traffic ones.

For high TIR , as the traffic is the main sound source, the similarity of the initial dictionary and \mathbf{W} is higher which allows retaining more elements as traffic components and then decreases the error (< 1 dB). The kept elements are then more suited to the scenes than a fixed dictionary. The error for these TIR is then due to the thresholding which put aside some elements that are nevertheless related to the traffic component. With a low threshold, it is possible to decrease the error. For $TIR = 12$ dB, the average error on all the sub-classes is $0.22 (\pm 0.08)$ dB with $t = 0.30$. In opposite for $TIR = -12$ dB, it is better to choose a high threshold $t = 0.55$, the error decrease then to $4.46 (\pm 1.66)$ dB. In order to generalize this method where no prior knowledge on the urban environment, the chosen threshold t is then fixed to $t = 0.42$ as it is the one that best balanced these opposite cases.

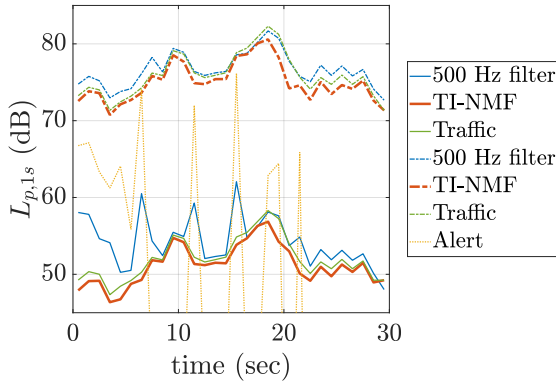


Fig. 10: 1 second equivalent sound pressure level of an *alert* sub-class scene for 500 Hz low-pass filter and TI-NMF ($\beta = 1$, $K = 200$, $w_t = 0.5$ s, $t = 0.42$) at $TIR = -12$ dB (in full line) and $TIR = 12$ dB (in dashed line).

5 Conclusion

In this work the non negative metric factorization framework was used to estimate the road traffic sound level in urban sound mixtures. It is a well suited approach to these sound environments because it easily takes into account the overlap between the multiple sound sources present in the cities and it is adapted to monophonic sensor networks. Different versions of NMF have been studied as a supervised and semi-supervised approach. On a large corpus of sounds, the supervised approach proves to be too restrictive to be adapted to different sound environments whereas the semi-supervised approach has, on the contrary, too many degrees of freedom on the mobile dictionary \mathbf{W}_r , decreasing its performance especially when the traffic is predominant. The proposed approach, named threshold initialized NMF achieves the lowest average error. With this method, the \mathbf{W} is initialized with road traffic spectra, updated and the dictionary elements that are similar to the road traffic spectra are then extracted by hard thresholding. A major advantage of the proposed approach is that it is not designed for a specific source. Even though the experiments described in this paper focused on traffic sounds, changing the dictionary to contain bird sounds would lead to an estimator of the presence of birds. Extending the approach to other sources is thus of interest for future research. Also, performance improvement could be achieved by the addition of constraints such as sparseness [49] and smoothness [31] of the low rank matrices.

The experimental protocol and the evaluated estimators have been implemented with the Matlab software. For reproducible purposes, the code is available

online³. The evaluation database composed of multiple samples of urban sounds is also made available for the research community with interest in detection, separation and recognition tasks of urban sound sources.

Acknowledgements The authors would like to acknowledge partial funding from the ANR under project reference ANR-16-CE22-0012.

References

1. H. Van Leeuwen, S. Van Banda, EuroNoise (2015). URL http://www.researchgate.net/publication/278406124_Noise_mapping_-_State_of_the_art_-_Is_it_just_as_simple_as_it_looks
2. O. Leroy, B. Gauvreau, F. Junker, E. De Rocquigny, M. Berengier, in *20th International Congress on Acoustics, ICA 2010* (France, 2010), p. 7p. URL <https://hal.archives-ouvertes.fr/hal-00636943>
3. N. Garg, S. Maji, Environmental Impact Assessment Review **46** (2014). DOI 10.1016/J.EIAR.2014.02.001. URL <https://www.osti.gov/scitech/biblio/22334150>
4. W. Wei, T.V. Renterghem, B.D. Coensel, D. Botteldooren, Applied Acoustics **Complete**(101), 127 (2016). DOI 10.1016/j.apacoust.2015.08.005. URL <https://www.infona.pl/resource/bwmeta1.element.elsevier-db3c0399-c7e0-343a-b534-d1757391efa2>
5. P. Mioduszewski, J.A. Ejsmont, J. Grabowski, D. Karpinski, Applied Acoustics **72**(8), 582 (2011)
6. C. Mietlicki, F. Mietlicki, M. Sineau, in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 2012 (Institute of Noise Control Engineering, 2012), vol. 2012, pp. 7119–7130
7. A. Can, T. Van Renterghem, D. Botteldooren, in *Acoustics 2012*, ed. by S.F. d'Acoustique (Nantes, France, 2012). URL <https://hal.archives-ouvertes.fr/hal-00811071>
8. D. Manvell, L. Ballarin Marcos, H. Stapelfeldt, R. Sanz, in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 2004 (Institute of Noise Control Engineering, 2004), vol. 2004, pp. 1998–2005
9. S. Xavier, S.J. Claudi, A. Francesc, B. Patrizia, al., Noise Mapping **3**(1) (2016). URL <https://doaj.org>
10. P. Bellucci, L. Peruzzi, G. Zambon, Applied Acoustics **Part B**(117), 193 (2017). DOI 10.1016/j.apacoust.2016.04.022. URL <https://www.infona.pl/resource/bwmeta1.element.elsevier-59e9924e-e72b-375e-8fce-41a09555f4e>
11. C. Mydlarz, J. Salamon, J.P. Bello, Applied Acoustics **117**, 207 (2017)
12. J. Salamon, J.P. Bello, IEEE Signal Processing Letters **24**(3), 279 (2017)
13. J. Picaut, A. Can, J. Ardouin, P. Crépeaux, T. Dhorne, D. Écotière, M. Lagrange, C. Lavandier, V. Mallet, C. Mietlicki, et al., The Journal of the Acoustical Society of America **141**(5), 3808 (2017)
14. T. Heittola, A. Mesaros, T. Virtanen, A. Eronen, in *in Workshop on Machine Listening in Multisource Environments, CHiME2011* (2011)
15. B. Defreville, F. Pachet, C. Rosin, P. Roy, (Audio Engineering Society, 2006). URL <http://www.aes.org/e-lib/browse.cfm?elib=13631>

³ <https://github.com/jean-remyGloaguen/article2017EstimationAmbiance>

16. A. Dufaux, L. Besacier, M. Ansorge, F. Pellandini, in *2000 10th European Signal Processing Conference* (2000), pp. 1–4
17. S. Chu, S. Narayanan, C.C.J. Kuo, *IEEE Transactions on Audio, Speech, and Language Processing* **17**(6), 1142 (2009). DOI 10.1109/TASL.2009.2017438
18. M. Cowling, R. Sitte, *Pattern Recognition Letters* **24**(15), 2895 (2003). DOI 10.1016/S0167-8655(03)00147-8. URL <http://www.sciencedirect.com/science/article/pii/S0167865503001478>
19. G. Shen, Q. Nguyen, J. Choi, *IFAC Proceedings Volumes* **45**(6), 1802 (2012). DOI 10.3182/20120523-3-RO-2023.00251. URL <http://www.sciencedirect.com/science/article/pii/S1474667016334127>
20. F. Beritelli, R. Grasso, in *2008 2nd International Conference on Signal Processing and Communication Systems* (2008), pp. 1–4. DOI 10.1109/ICSPCS.2008.4813723
21. L. Couvreur, M. Laniray, in *The 33rd International Congress and Exposition on Noise Control Engineering* (Prague, 2004)
22. J.C. Socoró, F. Aliás, R.M. Alsina-Pagès, *Sensors* **17**(10), 2323 (2017). DOI 10.3390/s17102323. URL <http://www.mdpi.com/1424-8220/17/10/2323>
23. H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, K. Shikano, *EURASIP Journal on Applied Signal Processing* **2003**, 1135 (2003)
24. D.D. Lee, H.S. Seung, *Nature* **401**(6755), 788 (1999). DOI 10.1038/44565. URL <http://www.nature.com/nature/journal/v401/n6755/abs/401788a0.html>
25. P. Smaragdis, J. Brown, in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on*. (2003), pp. 177–180. DOI 10.1109/ASPAA.2003.1285860
26. K.W. Wilson, B. Raj, P. Smaragdis, A. Divakaran, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (2008), pp. 4029–4032. DOI 10.1109/ICASSP.2008.4518538
27. A. Mesaros, T. Heittola, O. Dikmen, T. Virtanen, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), pp. 151–155. DOI 10.1109/ICASSP.2015.7177950
28. B. Wang, M.D. Plumbley, *Proc. DMRN summer conf* pp. 23–24 (2005). URL <http://www.eecs.qmul.ac.uk/~markp/2005/WangPlumbley05-dmrn.pdf>
29. I. Satoshi, K. Hiroyuki, *Computers & Mathematics with Applications* **64**(5), 1333 (2012). DOI 10.1016/j.camwa.2012.03.077
30. C. Févotte, N. Bertin, J. Durrieu, *Neural Computation* **21**(3), 793 (2009). DOI 10.1162/neco.2008.04-08-771
31. T. Virtanen, *IEEE Transactions on Audio, Speech, and Language Processing* **15**(3), 1066 (2007). DOI 10.1109/TASL.2006.885253
32. D. Lee, H. Seung, in *In NIPS* (MIT Press, 2000), pp. 556–562
33. A. Cichocki, R. Zdunek, in *Advances in Neural Networks – ISNN 2007* (Springer, Berlin, Heidelberg, 2007), Lecture Notes in Computer Science, pp. 793–802. DOI 10.1007/978-3-540-72395-0_97. URL https://link.springer.com/chapter/10.1007/978-3-540-72395-0_97
34. C.J. Lin, *Neural Computation* **19**(10), 2756 (2007). DOI 10.1162/neco.2007.19.10.2756
35. C. Févotte, J. Idier, *Neural Computation* **23**(9), 2421 (2011). URL http://www.mitpressjournals.org/doi/abs/10.1162/NECO_a_00168
36. H. Lee, J. Yoo, S. Choi, *IEEE Signal Processing Letters* **17**(1), 4 (2010). DOI 10.1109/LSP.2009.2027163
37. A. Lefevre, F. Bach, C. Févotte, in *ISMIR 2012: 13th International Society for Music Information Retrieval Conference* (2012)
38. C. Joder, F. Weninger, F. Eyben, D. Virette, B. Schuller, *Latent Variable Analysis and Signal Separation* pp. 322–329 (2012)
39. F. Weninger, J. Feliu, B. Schuller, in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* (IEEE, 2012), pp. 61–64
40. D.L. Donoho, *IEEE transactions on information theory* **41**(3), 613 (1995)
41. M. Fornasier, H. Rauhut, *Applied and Computational Harmonic Analysis* **25**(2), 187 (2008)
42. M. Rossignol, G. Lafay, M. Lagrange, N. Misdariis, in *1st Web Audio Conference (WAC)* (2015). URL <https://hal.archives-ouvertes.fr/hal-01078098/>
43. G. Lafay, M. Lagrange, M. Rossignol, E. Benetos, A. Roebel, *IEEE/ACM Transactions on Audio, Speech and Language Processing* **24**(10), 1854 (2016). DOI 10.1109/TASLP.2016.2587218. URL <https://hal.archives-ouvertes.fr/hal-01111381>
44. E. Benetos, G. Lafay, M. Lagrange, M.D. Plumbley, *IEEE Transactions on Audio, Speech and Language Processing* (2017). URL <https://hal.archives-ouvertes.fr/hal-01520194>
45. A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M.D. Plumbley, *IEEE/ACM Transactions on Audio, Speech and Language Processing* (2017). URL <https://hal.archives-ouvertes.fr/hal-01650601>
46. J. Salamon, C. Jacoby, J.P. Bello, in *22nd ACM International Conference on Multimedia (ACM-MM'14)* (Orlando, FL, USA, 2014)
47. J.R. Gloaguen, A. Can, M. Lagrange, J.F. Petiot, *The Journal of the Acoustical Society of America* **141**(5), 4044 (2017). DOI 10.1121/1.4989346. URL <http://asa.scitation.org/doi/abs/10.1121/1.4989346>
48. C.J. Willmott, K. Matsuura, *Climate research* **30**(1), 79 (2005)
49. P.O. Hoyer, *Journal of machine learning research* **5**(Nov), 1457 (2004)