Estimation of the road traffic sound levels based on the Semi-Supervised Non-Negative Matrix Factorization of Magnitude Spectrograms

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

Mathieu Lagrange
Jean-François Petiot
LS2N
Central School of Nantes

Abstract

1 Introduction

With the introduction of the European Directive 2002/EC/49, cities over 100 000 inhabitants have to produce road traffic noise maps. It allows Mathieu: who ? to estimate the number of city dwellers exposed to high noise levels and to draw up action plans to reduce it as too long exposures to these noises can generate health problems [1]. These maps are the result of a simulation process based on the estimation of the traffic density on the main roads and the use of sound propagation techniques. They express L_{DEN} and L_N, which are *Day-Evening-Night* and *Night* equivalent A-weighted sound levels respectively. However, these maps introduce lot of uncertainty generated by the numerical tools [2], by the different calculation methodologies used [3][4] or even by the calculation procedure of the number of inhabitants exposed to noise [5]. In addition, the usual road traffic noise maps are static, aggregating the exposure on the two indicators L_{DEN} and L_N, thus ignoring the sound levels evolution throughout the day. Since the creation of road traffic noise maps entails long data collection and calculation times, the use of acoustic measurements could facilitate their updating or even the generation of dynamic maps [6]. These measurements can be performed at fixed stations spread all over the cities [7] [8], which would lead to the availability of the long-term evolution of the traffic noise levels. It can also be performed with mobile stations [9] [10] covering a larger area with fewer sensors but also sparse time perdiods. The clustering between mobile and fixed measurements has been studied in [11]. Mathieu: je ne voispas l'intéret de cette dernière phrase

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

If sensors networks could improve road traffic noise estimation compared with simulated maps, the issue of the correct estimation from measurements of the traffic sound level is still unsolved [7]. Indeed, the urban sound environment is a complex environment gathering lots of different sounds (car passages, voices, bird's whistles, car horn ...) that can overlap. In consequence, the traffic sound level estimation based on measurements is not trivial task.

Many works have been 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 (Spectrum Gravity Spectrum, harmonicity, Mel-Frequency Cepstral Coefficient ...) and classify them with the help of classifiers (Support Vector Machines, Gaussian Mixture Models, Hidden Markov Model, Artifical Neural Networks). A description of there features and classifiers can be found in [18] and their application can be found in [19], [20], [21]. However, most of these results in the detection or recognition tasks, do not address the overlap of environmental sounds in an urban context. Although, near major roads or ring roads, traffic is predominant on all other sound sources, there are many places where road traffic overlaps with other sound sources that contribute significantly to the overall sound levels. In such case, the only detection of the traffic component does not make it possible to determine precisely its noise level. In consequence, to be effective on all the different sound environments, it seems more suited to consider the issue as the one of blind source separation.

One of the first and the most widely used techniques

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

is the Independent Component Analysis [22]. The principle is to decompose N recorded signals to a sum of P independent sound sources weighted by linear relations. This method is most of all suited for the 'cocktail party' issue where one tries to capture a signal among noise. However, ICA is limited to only over determined cases ($N > P$). Furthermore, if it is suited for indoor environments where the number of sound sources is constant, it can not be fitted for an outdoor environment where the number of sources is unknown and variable and, moreover, it would be necessary to install multiples sensors on one point to realize the source separation which is not feasible. A second method is Non-negative Matrix Factorization (NMF) [23] which consists in approximating the non-negative spectrogram of an audio file from the product of two matrices. It has been widely used in the audio domain, [24] [25] [26], and has already been employed for the source separation task of monaural signals of speech and music [27] [25]. This method has the advantage to easily deal with the overlap of the sound sources. For the environmental sounds, the method has been used for the geo-localisation and classification of the sound environment, like in [28] where NMF allows to classify the audio files according to the 10 cities where they have been recorded. For the source separation, it has been used by Innami and Kasai in the unsupervised case [29]. They proposed a source separation in two steps by separating the sound background from the events first and by separating the events between them. The audio files tested results of a simulation process where a sound background (river or wind) are adding to two sound events (school chime, announcement, frog croaking, dog barging and bell ringing). If the method proposed is interesting, the main issue here is the small size of the database (only 9 sounds) on which the algorithms are tested while some sounds (frog and river) are not representative of sounds that can be found in cities.

Our study proposed to applied Non-Negative Matrix Factorization on simulated sound scenes to estimate the traffic sound level. The use of simulated sound scenes is necessary as it offers a full control on the design of the scenes and the knowledge of the exact contribution of the traffic component (what recordings do not allow). Part 2 details the technical aspect of NMF. Part 3 described on the experimental protocol set up. Then part 4 reveals the results obtained during the parametric study.

2 Non-negative Matrix Factorization

2.1 Description of NMF

Non-negative Matrix Factorization (NMF) is an approximation method introduced by Lee and Seung, [23], which estimates the spectrogram (get from a Short-Term Fourier Transform) of an audio file, $\mathbf{V}$, $\in \mathbb{R}^{+}_{F\times N}$ as :

$$\mathbf{V} = \tilde{\mathbf{V}} \approx \mathbf{WH} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{+}_{F\times K}$ is the *dictionary* (or basis) matrix composed of audio spectrum and $\mathbf{H} \in \mathbb{R}^{+}_{K\times N}$ is the *activation* matrix which summarizes the temporal evolution of each element of $\mathbf{W}$ (fig. 1).

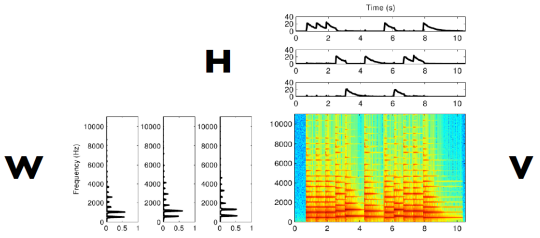

Figure 1: Example of a simple NMF for a musical content [30]

The choice of the dimensions is imposes such as $F \times K + K \times N < F \times N$. To estimate the quality of the approximation, an objective function is used

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

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

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

and usually belongs to the $\beta-$divergence class [31] in which the well known Euclidean distance (eq. 4a) and the Kullback-Leibler divergence (eq. 4b) belong

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

The prior knowledges on the content can be adjusted with the add of constraints (like the smoothness or the sparsness criteria [32]) in the objective function (equation (2)) to better take account prior knowledge of the sources.

Algorithms have been proposed to solve the minimization problem (2) iteratively such as the multiplicative upatde, the alternating least square method [33], the projected gradient [34] ... Here, the multiplicative update is chosen [35] as it ensure non-negative results of which convergence has been proved [36].

2.2 Supervised NMF

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

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

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 $A.B$ and A/B symbolized the Hadamard product and ratio. The source separation is made by extracting the dictionary and basis elements related to the traffic

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

2.3 Semi-supervised NMF

One of the main issue with the supervised approach is this is not always adapted to different sound environments as the fixed dictionary has to sum up all the sound sources that can be present. In consequence, as the dictionary dimension is limited and cannot include all the urban sounds, to offer more flexibility, semi-supervised NMF has been proposed [37]. This method consists in composing the *dictionary* with a fixed part $\mathbf{W_s} \in \mathbb{R}^+_{F \times K}$ (composed here of traffic audio spectrum) and with a mobile part, $\mathbf{W_r} \in \mathbb{R}^+_{F \times J}$ with $J << K$, that is updated. Here, $J = 2$. The aim is to include in $\mathbf{W_r}$ the element that are not related with the traffic. The problem (1) become

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

In a similar way as to solve the equation 2, $\mathbf{W_r}$, $\mathbf{H_r}$ and $\mathbf{H_s}$ are successively updated with the relations (8):

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

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

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

3 Experimental protocol

Simulated sound scenes are used to assess the performance of NMF. This offers a controlled framework to design specific sound environments in which all the traffic component is known. Then, the road traffic sound levels estimated with the method can be compared to the real ones, introduced within each simulated sound scene.

3.1 Environmental sound scene corpus

A corpus is designed with the *simScene* software[2]. *simScene* [38] is a simulator that creates sound scenes in a .wav format by superposing audio samples that come from an isolated sound database. This database is divided in two categories: *i*) the *event* category which are the brief sounds (from 1 to 20 seconds) that are considered as salient, *ii*) the *background* category includes all the sounds that are of long duration and whose acoustic properties do not vary with respect to time. Inside each category, the sound samples are grouped in sound classes (*bird, car, foot steps* ...), each of them being composed of multiples samples (bird01.wav, bird02.wav ...).

The software allows the user to control some parameters (number of events of each class that appear in the mixture, elapsed time between each sample of a same class, presence of a fade in and a fade out ...) completed with a standard deviation that may brings some random behavior between the scenes. Furthermore, an audio file of each sound class present in the scene can be generated that allows to know its exact contribution as well as a text file that summarizes the time presence of all the events.

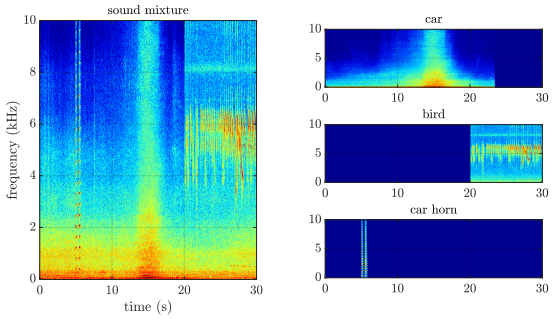

Figure 2: Example of a sound scene composed of 3 sound classes (car, bird, car horn)

A sound database has been built including 245 sound event samples divided in 19 sound classes (*ringing bell,*

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

birds, sweeping broom, car horn, car passages, hammer, drill, coughing, barking dog, rolling suitcase, closing door, plane, siren, footstep, storm, street noise, train, tramway, truck and voice) and 154 sound background samples divided in 9 sound classes (*birds, construction site noise, crowd, park, rain, children playing in schoolyard, constant traffic noise, ventilation, wind*). The sound class *car passages* comes from recordings of 4 cars made on the Ifsttar's runway on different speeds with multiple gear ratio. The other audio files have been found online (*freesound.org*) and with the help of the *UrbanSound8k* database [39]. This database enables creating realistic urban sound scenes from the road traffic point of view [40]. A sound mixing corpus is composed of 6 sub-corpuses of 25 audio files each lasting 30 seconds. Each sub-corpus is characterized by a specific generic sound class : *alert* (car horn, siren), *animals* (barking dog , whistling birds), *climate* (wind, rain), *humans* (crowd noise and voice), *mechanics* (different metallic and construction site noise) and *transportation* (train, tramway and plane). In each file, traffic component is present as the sum of the background and event traffic sounds and is mixed with the sound classes. The sound classes that are not related to the traffic component are summed up as the *perturbator* (n'existe pas comme terme, disruptive à la place ?) class. To test different scenarios, each audio file is duplicated with the traffic sound level of the entire sound scene, $L_{p,traffic}$, fixed to a specific level according to the sound level of the *perturbator* class, $L_{p,perturbator}$ following the relation (9).

$$TPR = L_{p,traffic} - L_{p,perturbator} \tag{9}$$

with the *Traffic Perturbator Ratio* $TPR = [-12, -6, 0, 6, 12]$. When $TPR = -12$, the traffic component is then less present than when $TPR = 12$ where it is predominant on the *perturbator* class. The 1 second equivalent sound pressure level, $p_{1s,traffic}$, is also calculated (figure 3). Finally, the number of scenes designed is 750 (6 sub-corpus $\times$ 25 scenes $\times$ 5 TPR values).

3.2 Experiment

The experiment consists in estimating the traffic road sound level on the environmental sound corpus. NMF is used and compared to a simple approach which consist in a frequency low-pass filter (this second method relies on the fact that road traffic is mainly composed of low frequencies, compared to main of the confounded sounds (voices, birds, etc.)) with the cut-off frequency f_c. The spectrogram $\mathbf{V}$ is built with a window size $w = 2^{14}$ with a 75 % overlapping and a number of point $nfft = 2^{14}$. Therefore, the dimensions of V are $F = 8192$ and $N = 772$. Similarly, the frequency low-pass filter is applied on the spectrogram $\mathbf{V}$ and on the dictionary $\mathbf{W}$ (renamed $\mathbf{V}_{f_c}$ and $\mathbf{W}_{f_c}$ respectively). The aim to focus the reconstruction of the signal on the frequencies where the traffic components are. The figure 4 summarizes the different steps of the process depending on the chosen method, which consist of 3 steps:

- the dictionary building for NMF,
- the estimation of the traffic sound level according the chosen method (NMF or frequency low-pass filter),
- the calculation of the error between the exact and the estimated sound levels.

3.2.1 Dictionary building

The dictionary is built from a sound database dedicated specially to this task. It is composed of 53 audio files of car passages. First, for each audio file, its spectrogram is calculated with fixed parameters (w, 75 % overlap, $nfft$). Then a temporal rectangular window is applied without overlapping on the spectrogram in order to consider several spectrum for each audio file. The size of the window is set up at 0.5 and 1 second. In each window, the root mean square value is calculated on each frequency bin to reduce the different spectrum in one spectra. An example that illustrates the process can be found on figure 5 in the ease of a large window of 4 seconds.

However, the number of elements got on all the sound database does not respect the constraint imposed about the dimension of NMF ($F \times K + K \times N < F \times N$). With a 1 second window, $K = 1003$. In consequence, a K-medoid clustering is applied to reduce the number of spectrum to $K = [25, 50, 100]$. A special case is added where the root mean square of *all* the spectrogram is applied. One spectra is then generated by audio file. Each element is normalized such as $||\mathbf{W_k}|| = 1$ with $||\bullet||$ is the $\ell - 1$ norm. Table 1 summarizes the parameters and their related values.

K	25	50	100
temporal window (s)	0.5	1	*all*

Table 1: Summary of the dictionary parameters

3.2.2 Estimation of the traffic sound level

The traffic sound level is estimated according to the choice of the estimator (a frequency low-pass filter or

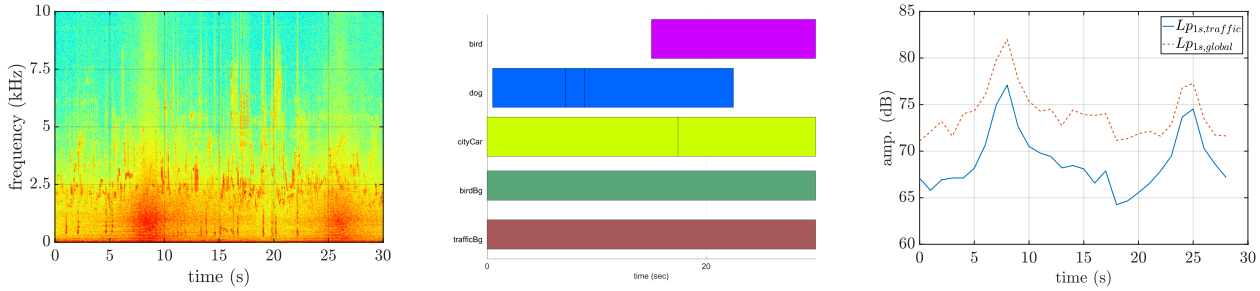

Figure 3: Example of a scene of the *animals* sub corpus. Spectrogram (on left), *Piano Roll* of the different sound classes (on the middle) and 1-s equivalent sound level of the traffic, $Lp_{1s,traffic}$ and of the global sound scene, $Lp_{1s,global}$(on right)

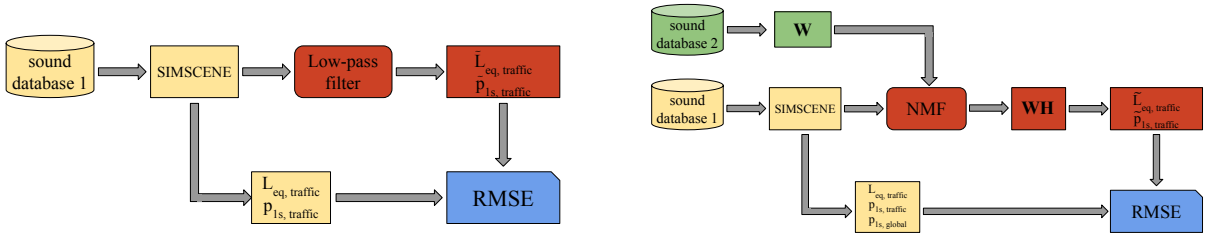

Figure 4: Block diagrams summed up the different step of the precess. On top, for the frequency low-pass filter, on bottom, for NMF.

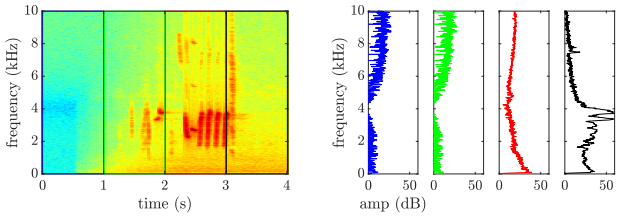

Figure 5: Example of the extraction of the spectrum with a 1 second window. On left, the original spectrogram with the temporal window, on the left, the 4 spectrum obtained (dB scale)

NMF). These methods are applied on the totality of the corpus (*alert* (al), *animals* (an), *humans* (hu), *climate* (cl), *mechanics* (me), *transportation* (tr)), for all the cut-off frequencies f_c ([500 1k 2k 5k 10k 20k] Hz) and for 5 TPR ([-12 -6 0 6 12] dB).

3.2.2.1 Frequency low-pass filter

The first estimator to determine the traffic sound level is a basic frequency low-pass filter which depend only on the cut-off frequencies f_c.

3.2.2.2 NMF

The second estimator is supervised and semi-supervised NMF (see part 2). The spectrogram $\mathbf{V}$ and the dictionary $\mathbf{W}$ are expressed on two different formats: with a linear frequency scale ($\Delta f \approx 2.8$ Hz) and with third octave bands (27 bands). Theses two methods enable to compare a fine approach (the linear scale) with a degraded one (the third octave bands) as it reduces the number of frequency bins and allows to reduces the number of bands in the high frequencies where the traffic component is less present. NMF is performed on 400 iterations which is sufficiently enough to get a stabilized reconstruction. Furthermore, in the case of the linear frequency scale, $\mathbf{V}$ and $\mathbf{W}$ are filtered at the frequencies f_c in order to focused the reconstruction of the signal of the low frequency bins. In consequence, if NMF is performed with filtered elements ($\mathbf{V}_{f_c}$ and $\mathbf{W}_{f_c}$) to determine $\mathbf{H}_{f_c}$, the traffic signal reconstruction is made with the original dictionary $\mathbf{W}$, as

$$\tilde{\mathbf{V}}_{traffic} = [\mathbf{W}\mathbf{H}_{f_c}]_{traffic} \tag{10}$$

For the third octave frequency scale, only the case $f_c = 20$ kHz is applied. During the iteration process, the estimated equivalent traffic sound level in dB of the entire scene, $\tilde{L}_{p,traffic}$, is calculated as well as the 1 second equivalent sound pressure level, $\tilde{p}_{1s,traffic}$. Table 2 summarizes the parameters and their related values.

f_c (kHz)	0.5	1	2	5	10	20
TPR (dB)	-12	-6	0	6	12	
sub-classes	alert	animals	climate	humans	transportation	mechanics
method	filter		supervised NMF		semi-supervised NMF	
β	1				2	
frequency scale	linear				third octave	

Table 2: Summary of the different parameters and their values for the estimation of the traffic sound level

3.2.3 Metric

The performances of the two estimators of the road traffic sound level are assessed through the calculation of two metrics.

- The Mean Absolute Error, MAE, expresses the quality of the long-term reconstruction of the signal. It consists in the average over the N sound scenes of the absolute difference between the exact and estimated traffic sound level in dB,

$$MAE = \frac{\sum_{n=1}^{N} |L^n_{p,traffic} - \tilde{L}^n_{p,traffic}|}{N}. \quad (11)$$

- The normalized short-term Root Mean Square Error, $nRMSE$, calculates, for each sound scene, the error between the exact and estimated road traffic 1-s equivalent sound level of each file normalized by the 1-s equivalent sound level of the global scene, $p^t_{1s,global}$, in the linear scale,

$$nRMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{p^t_{1s,traffic} - \tilde{p}^t_{1s,traffic}}{p^t_{1s,global}}\right)^2} \quad (12)$$

with T is the number of temporal bin in the signal. The linear scale is here more relevant as it is more sensitive to the error on the high sound levels than the dB scale. Then for one combination of factors, the N $nRMSE$ calculated are averaged.

In all, according the table 1 and 2, 6848 settings are performed between the different dictionary **W** and the multiple parameters of the estimation step.

4 Results

4.1 Frequency low-pass filter results

In a first time, the frequency low-pass filter estimator is performed on all the scene. The table 3 summarized the mean error on the totality of the 750 scenes to find the most efficient cut-off frequency.

f_c (Hz)	MAE (dB)	$nRMSE$
500	**5.66 ±6.59**	**1.48 ±1.10**
1000	**6.31 ±7.64**	**1.52 ±1.34**
2000	**6.65 ±8.24**	**1.57 ±1.54**
5000	**7.42 ±8.90**	1.82 ±1.65
10000	**7.55 ±9.00**	1.89 ±1.70
20000	**7.59 ±9.02**	1.89 ±1.72

Table 3: RMSE error for the low pass filter averaged on all the TPR and sub-classes

According the table 3, the cut-off frequency at 500 Hz is the most efficient on all the TPR and all the sub-classes. The corresponding errors according to the sub-classes and the TPR are summarized in the figure 6.

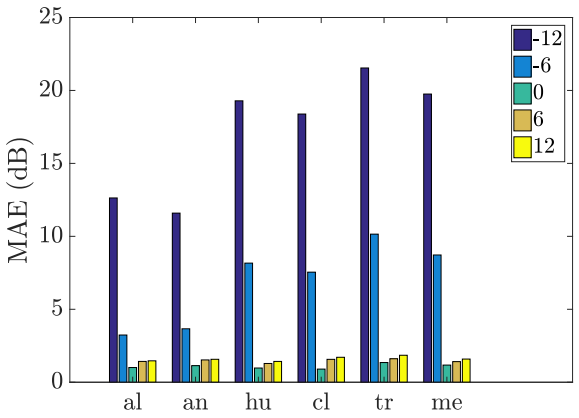

Figure 6: Bar plot for the filter according to the sub-classes and the TPR at $f_c = 500$ Hz cut-off frequency

The error for all the sub-corpus is important for the low TPR (-12 and -6) due to the confusion between the *perturbator* class and the *traffic* class. The error is less for the *alert* and *humans* sub-corpus as it is composed of higher frequencies while, for other sub-classes, the error are far more important. For the higher *TPR*, the traffic is more present and become dominating on the *perturbator* class. The lower error, here, is due to the suppression of the traffic energy by the filter. Finally, the error for $f_c = 20$ kHz is equivalent to the one produced when the *traffic* and the *perturbator* classes are take into account together without distinction.

4.2 Supervised NMF results

As it is not possible to summarize all the parameter combinations, just the best results according to the domain and β at the $400th$ iteration are presented in table 4. The bar plot in figure 7 displays the MAE error for the best combination.

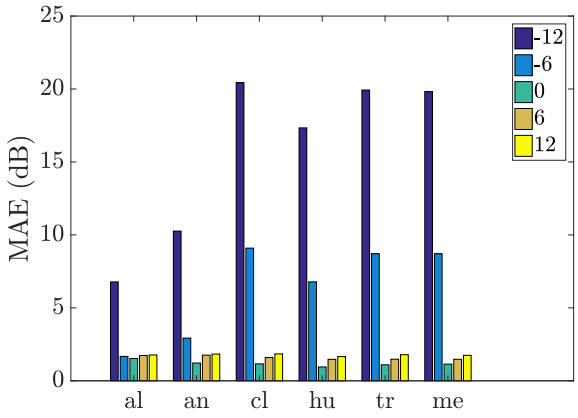

Figure 7: bar plot for the best parameter combination of the supervised NMF ($K = 25$, temporal window = 0.5 s, domain = spectra, $\beta = 2$, $f_c = 500$ Hz)

The best combination for supervised NMF is the one with described in the *spectra* domain and for the euclidean distance with $K = 25$, a temporal window of 500 ms and $f_c = 500$ Hz. The errors produced by NMF, on the totality of the scenes, is smaller than the filter approach for all the metrics. For the higher *TPR*, both methods have similar performances (table 5). For the lower *TPR*, TPR, supervised NMF reduces significantly the error. By distinguishing each class, it is lower for the *alert* and *animals* than the rest sub-classes.

This difference can be illustrate through the evolution of the cost function (figures 9) and of the MAE error (figure 8) for 3 sub-classes (*alert*, *climate* and *transportation*), for 2 *TPR* (-6 and 6) until the $100th$ iterations.

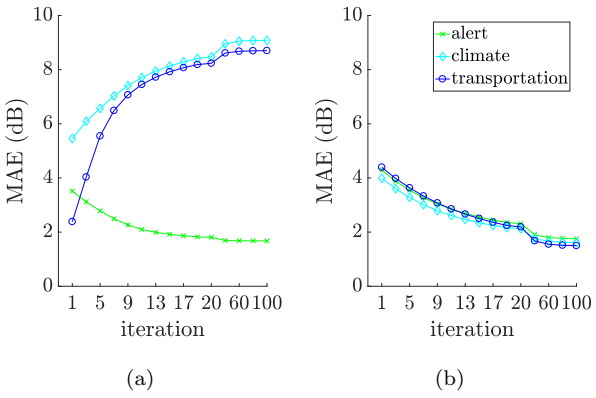

Figure 8: Evolution of the MAE for 3 sub-classes for TPR = -6 (fig. 8(a)) and TPR = 6 (fig. 8(b))

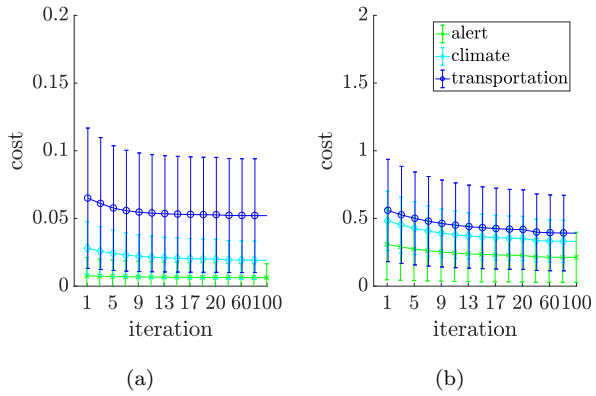

Figure 9: Evolution of the MAE for 3 sub-classes for TPR = -6 (fig. 9(a)) and TPR = 6 (fig. 9(b))

For the TPR = 6, the cost function and the MAE error are decreasing, meaning the global mixture and the traffic signal are well synthesized. Here as the $\mathbf{W}$ is only composed of traffic elements and the sound scenes are composed of a predominant traffic, the reconstruction of the signal is easier. In opposite, for the case where TPR = -6, for the *alert* signal, the cost function and the MAE error are decreasing too. But for the *climate* and *transportation* sub-classes, the MAE error is increasing. Even if the cost function decreases, the quality of the traffic signal rebuilt is not improved. This difference can be explained by the type of sound present in these different sub-classes: for *alert* and *animals*, it includes harmonic sounds which belong in the frequency range $[2000 - 5000]$ Hz while the other sub-classes include lower frequency sounds. As $\mathbf{W}$ is composed of traffic spectrum, located in low frequencies too, it is more

K	temporal window (ms)	f_c (Hz)	domain	β	MAE (dB)	$nRMSE$
25	500	spectra	2000	1	**6.18 ±7.49**	**1.48 ±1.20**
25	500	spectra	500	2	**5.32 ±6.28**	**1.40 ±0.93**
25	0	third octave	20000	1	**6.82 ±8.16**	1.60 ±1.34
25	500	third octave	20000	2	**6.26 ±7.69**	**1.36 ±1.19**

Table 4: RMSE error for supervised NMF averaged on all the TPR and sub-classes

	MAE (dB)	
TPR	filtre	supervised NMF
-12	17.20 ± 4.09	15.76 ± 5.82
-6	6.91 ± 2.82	6.32 ± 3.24
0	1.09 ± 0.16	1.18 ± 0.19
6	1.47 ± 0.12	1.59 ± 0.13
12	1.60 ± 0.16	1.77 ± 0.07

Table 5: Comparison of the MAE error for the filter at 500 Hz and the best combinaison of the supervised NMF

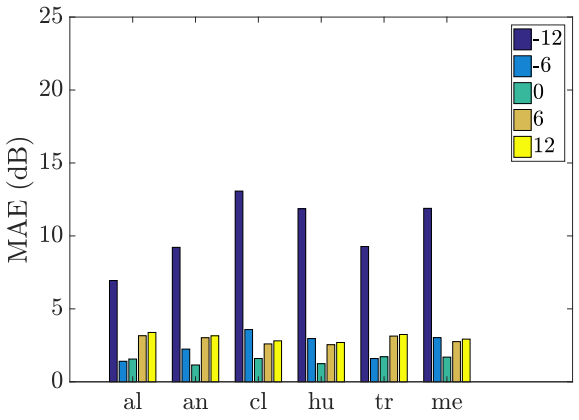

Figure 10: bar plot for the best parameter combination of the supervised NMF (K = 25, temporal window = 0.5 s, domain = spectra, β = 2, f_c = 20 kHz)

difficult to recompose correctly the traffic signal when the *perturbator* sound is predominant and composed of low frequencies too. As NMF minimized the cost function (equation 2), *traffic* spectrum then are used to synthesized the global mixture at the expense of the traffic signal.

4.3 Semi-supervised NMF results

The errors produced for the semi-supervised approach is summarized in table 6 according to β and the frequency domain (spectral or third octave).

The best combination is got in the third octave domain with $\beta = 1$, $f_c = 20$ kHz, $K = 25$ and a temporal window null. The error, compared to the supervised approach, is lower. Furthermore, the standard deviation is lower too meaning that the semi-supervised approach offer a much more stable error than the previous method. The figure 10 displayed the error for the best scenario averaged on all the sub-classes and the TPR. The table 7 summarizes the error expanded to the TPR.

The error for low TPR is much lower than the filter or supervised NMF. The add of the mobile part makes NMF less constraint and therefore allows to adapt NMF to low traffic contents. Figure display the mobile part of the dictionary, $\mathbf{W_r}$ for two particular sub-classes (*alert* and *climate* and for TPR = -6.

figure W_r for alert and climate

Nevertheless, the advantage won with the mobile part, generates larger errors in higher TPR. Here, NMF uses $\mathbf{W_r}$ in order to minimize the cost function. As the mobile part is not constraint, traffic content is included in it, decreasing the quality of the traffic signal reconstruction. Figure display the 2 basis of $\mathbf{W_r}$.

figure W_r for alert and climate

These behaviors can be noticed through the MAE evolution in figures 11: for low TPR the error is decreasing like the cost function (figure 12(a)) meaning that the synthesis of the traffic signal is good. For the *alert* sub-class, $\mathbf{W_r}$ enables to include the harmonic component which are not present in $\mathbf{W_s}$, the improvement bring by the semi-supervised approach is then remarkable. Whereas, in *climate* sub-class, the *perturbator* sound classes are composed of low-frequencies that are present in $\mathbf{W_s}$. This class might be describe with the help of $\mathbf{W_s}$.

For TPR = 6, the error is increasing for all sub-classes. As the objective function is to minimize the distance between the spectrogram $\mathbf{V}$ and $\mathbf{WH}$, the semi-supervised approach is free to include traffic component in $\mathbf{W_r}$ to do it. This degree of freedom deteriorates the reconstruction of the traffic signal.

miss the evolution of the Lp for the 25 scenes,

K	temporal window (ms)	f_c (Hz)	domain	β	MAE (dB)	$nRMSE$
50	0	spectra	1000	1	**4.50 ±3.05**	1.45 ±0.76
100	0	spectra	20000	2	**4.05 ±3.43**	**1.16 ±0.53**
25	0	third octave	20000	1	**3.96 ±2.18**	1.34 ±0.72
25	0	third octave	20000	2	**4.29 ±3.58**	**1.21 ±0.61**

Table 6: RMSE error for semi-supervised NMF averaged on all the TPR and sub-classes

	MAE (dB)	
TPR	filter	semi-Supervised NMF
-12	17.20 ±4.09	7.46 ±2.88
-6	6.91 ±2.82	2.15 ±0.70
0	1.09 ±0.16	2.20 ±0.45
6	1.47 ±0.12	3.61 ±0.29
12	1.60 ±0.16	4.02 ±0.33

Table 7: Comparison of the baseline and semi-supervised NMF with the best combination

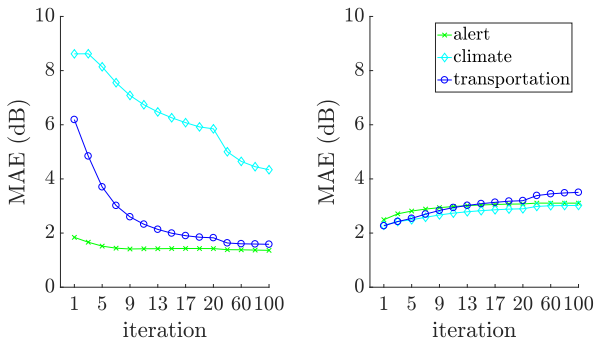

Figure 11: Evolution of the MAE for 3 sub-classes for TPR = -6 (fig. 11(a)) and TPR = 6 (fig. 11(b))

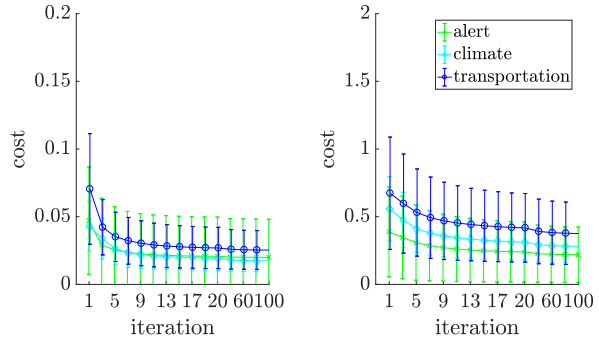

Figure 12: Evolution of the MAE for 3 sub-classes for TPR = -6 (fig. 12(a)) and TPR = 6 (fig. 12(b))

estimate and exact + 2/3 scenes to see the evolution during the 30 seconds of the scenes ?

5 Conclusion

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[17] S. Chu, S. Narayanan, and C. C. J. Kuo. Environmental Sound Recognition With Time-Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, August 2009.

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

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

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

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

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

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

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

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

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

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

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

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

[30] N. Bertin. *Les factorisations en matrices non-négatives : approches contraintes et probabilistes, application á la transcription automatique de musique polyphonique.* Paris, Télécom ParisTech, January 2009.

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

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

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

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

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

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

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

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

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

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