Estimation of the road traffic sound levels based on Non-Negative Matrix Factorization

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

1.1 Related work

With the introduction of the European Directive 2002/EC/49, cities over 100 000 inhabitants have to produce road traffic noise maps. Theses maps allow 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 periods.

Currently, sensor networks in cities are spread for multiple applications (air quality assessment, measurement of meteorological parameters, ...), including the assessment of urban noise levels. DYNAMAP project [11] studied the establishment and feasibility of such installations. It focuses on sensor installations on specific roads at the city scale in Milan and Rome [12]. 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. Consequently, the traffic sound level estimation based on measurements is not trivial task. Many recent works have focused on the detection or recognition tasks of environmental sounds without distinction between them[13], [14], [15], [16]. 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 [17] and their application can be found in [18], [19], [20]. Recently, an Anomalous Noise Events Detector has been generated in [21] to detect the sound sources from labeled recordings that are not related to the traffic component in order not to take them into account on the estimation of the traffic sound level. If the detection of the road traffic noise is good, the detection of theses anomalous noise events stay weak and no information on the improvement on the estimation of the traffic nous level are presented.

Furthermore, this work and as well as the other works 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

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

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 a wide range of sound environments, we propose in this paper to follow the blind source separation paradigm. That is, separating the contribution of the traffic from the other sources within a polyphonic scene.

One of the first and the most widely used techniques to do so 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 mount multiples sensors on one point to perform the source separation. A more convenient method is Non-negative Matrix Factorization (NMF) [23] which consists in approximating the magnitude 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]. By design, this method deals with the overlaping sound sources as soon as the overlap can be resolved on the time/frequency plane. For the environmental sounds, the method has been used for the geo-localisation and classification of the sound environment, like in [28] where NMF is used to classify the audio files according to the 10 cities where they have been recorded. It has also been used by Innami and Kasai in the unsupervised case [29] for source separation. 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.

1.2 Proposed approach

We propose in this paper a method based on the Non-Negative Matrix Factorization (NMF) technique to estimate the global, $\tilde{L}_{p,traffic}$, and the 1-s equivalent, $\tilde{p}_{1s,traffic}$ sound level of the traffic through the supervised and the semi-supervised approaches. To validate this approach, we consider a corpus of simulated scenes artificially created with the simulator software *simScene*. 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 which would hardly be extracted from a recording of an urban scene ($L_{p,traffic}$ and $p_{1s,traffic}$). Both the sound scene simulation and NMF require the creation of two sound databases (figure **??**). In parallel, a baseline method built from a frequency low-pass filter is computed. This method considers that road traffic is mainly composed of low frequencies and therefore can be filtered by a low-pass filter at the cut-off frequency f_c (figure **??**). The performance of the frequency low-pass filter and NMF are then compared with the calculation of two metrics (Mean Absolute Error, normalized Root Mean Square Error).

The remaining of the paper is organized as follows. Section 2 details the technical aspect of NMF. Section 3 described on the design of the environmental sound scene corpus and the experimental protocol setup. Then Section 4 shows and discusses the results obtained during the parametric study.

2 Non-negative Matrix Factorization

2.1 Description of NMF

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

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

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).

The choice of the dimensions is often made as 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 essential. 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:

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

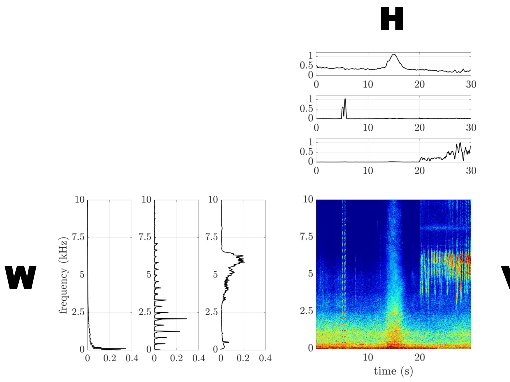


Figure 1: Example of a simple NMF for urban sound mixture, **W** and **V** are composed of 3 elements (car passages, car horn and bird's whitles)

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

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

Prior knowledge on the content can be adjusted with the addition of constraints (like the smoothness or the sparsness criteria [31]) 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 update [32], the alternating least square method [33], the projected gradient [34] ...Here, the multiplicative update is chosen as it ensure non-negative results of which convergence has been proved [35].

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 estimate. In the first iteration, **H** is initialized randomly, then it is updated by the generic algorithm

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

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 symbolized the Hadamard product and ratio. As in the supervised approach, the position in **W** of traffic component is known, the source separation of this sound source is made by extracting the dictionary and basis elements related,

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

2.3 Semi-supervised NMF

One of the main issue with the supervised approach is the generalization issue: how to be adapted to different sound mixtures with a fixed dictionary ? To better take account for the diverse nature of urban scenes, semi-supervised NMF can be useful as it has been proposed [36] to offer more flexibility. This method consists in composing the *dictionary* with a fixed part $\mathbf{W_s} \in \mathbb{R}^+_{F\times K}$, composed in our case of spectrum representative of road traffic 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

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

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):

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

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

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

2.4 Thresholded constrained NMF

A last approach is tested based on unsupervised NMF. Usually, **W** is learnt with the help of a learning corpus by initiated it randomly. Here, as the concerned sound source is known and audio samples of car passages are available, a 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 **W** (eq. 9) and **H** (eq. 5) are updated alternatively.

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

After N iterations, a measure of similarity $D_\theta\left(\mathbf{W_0}||\mathbf{W}\right)$ between $\mathbf{W_0}$ and the get dictionary $\mathbf{W}$ for each element k is computed through a cosine similarity,

$$\cos\theta = \frac{\mathbf{W}.\mathbf{W_0}}{||\mathbf{W}||.||\mathbf{W_0}||}. \tag{10}$$

$\cos\theta = 1$ means that the elements are identical (the k-th element of $\mathbf{W}$ is then considered as traffic element) whereas $\cos\theta = 0$ means that the elements are significantly different. This measure allows a bound between 1 and 0 and is an invariant scale estimation of the similarity. Then, the similarity is sorted in descending order. The elements in $\mathbf{W}$ that can belong to $\mathbf{W}_{traffic}$ are then selected by a *hard thresholding* method (Figure 2). It is defined as :

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

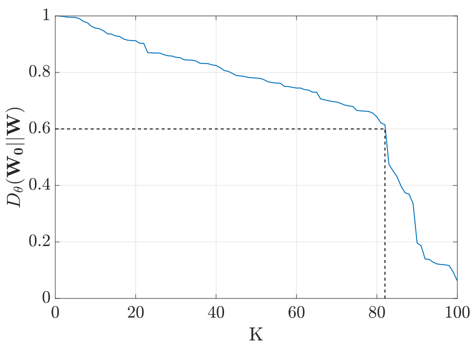


Figure 2: Example of the $\mathbf{W}_{traffic}$ extraction from the sorted cosine similarity with a threshold $t = 0.6$. The 82-nd first elements are considered as traffic component.

Other thresholding methods as the *soft* [37]and the *firm* [38] and multiples way to display the distance through a sigmoïd or a Radial Basis Function have been investigated. A fast parametric study has revealed that the *hard* tresholding method with a linear representation of the similarity according to K (see fig. 2) was the best way to get better performances.

3 Experimental protocol

In order to validate the usefulness of considering NMF framework to estimate the road traffic noise level, one need to have a reference level. It can hardly be measured or even annotated from real life recordings. Thus, simulated sound scenes are used to assess the performance of the proposed 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* [39] is a simulator that creates sound scenes in a .wav format by summing 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 including 245 sound event samples divided in 19 sound classes (*ringing bell, 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 *ii*) the *background* 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 belong to this category 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 60 recordings of 2 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 within the *UrbanSound8k* database [40]. Each sound classes are 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.

This database enables creating realistic urban sound scenes from the road traffic point of view [41]. 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

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

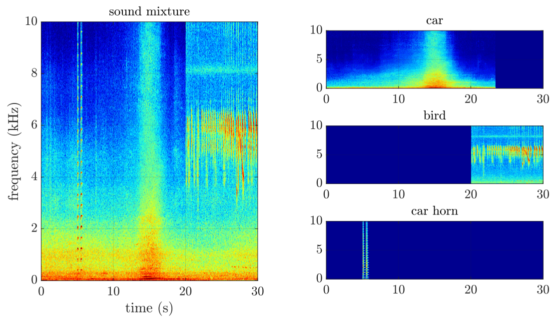


Figure 3: Example of a sound scene composed of 3 sound classes (car, bird, car horn)

with traffic will make the estimation of the traffic level more difficult. The classes are : *alert* (car horn, siren), *animals* (barking dog , whistling birds), *climate* (wind, rain), *humans* (crowd noise and voice), *mechanics* (different 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 and is mixed with the sound classes. The sound classes that are not related to the traffic component are summed up as the *interfering* sound 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 *interfering* class, $L_{p,interfering}$ following the relation (12).

$$TIR = L_{p,traffic} - L_{p,interfering} \tag{12}$$

with the *Traffic Interference Ratio* $TIR = [-12, -6, 0, 6, 12]$. When $TIR = -12$, the traffic component is then less present than when $TIR = 12$ where it is predominant on the *interfering* class. The 1 second equivalent sound pressure level, $p_{1s,traffic}$, is also calculated (figure 4). 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 of the 6 environmental sound sub-corpus (*alert* (al), *animals* (an), *humans* (hu), *climate* (cl), *mechanics* (me), *transportation* (tr)) and for 5 TIR ([-12 -6 0 6 12] dB). First, the spectrogram $\mathbf{V}$ of each sound scene is built with a window size $w = 2^{12}$ with a 50 % overlap and a number of point $nfft = 2^{12}$. Therefore, the dimensions of $\mathbf{V}$ are $F = 2049$ and $N = 664$.

The first estimator to determine the traffic sound level is the basic frequency low-pass filter which depend only on the cut-off frequencies f_c = [500 1k 2k 5k 10k 20k] Hz. The spectrograms $\mathbf{V}$ are filtered and the remaining energy is then considered as traffic component (eq. 13).

$$\tilde{\mathbf{V}}_{traffic} = \mathbf{V}_{f_c}. \tag{13}$$

The second estimator is NMF. Multiples parameters are involved here between the building of the dictionary and the performed NMF.

3.2.1 Dictionary building

The dictionary is built from a second sound database dedicated specifically to this task. It is composed of 53 audio files of passing cars. These records have been made on the Ifsttar's runway with the same conditions that the records made for the *SimScene* database but with two different cars (). First, for each audio file, its spectrogram is calculated with fixed parameters (w, 50 % overlap, $nfft$). Then time/frequency windows with $w_t \times F$ dimension are applied without overlapping on the spectrogram in order to consider several spectrum for each audio file. w_t is fixed at $wt = [0.51]$ second. In each window, the root mean square value is calculated on each frequency bin to reduce the different spectrum in one spectra. Since the number of elements given by processing all the sound database is high, in order to reduce the computational and delete redundant information, a K-means 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. Each audio file generates one element k of $\mathbf{W}$. An example that illustrates the process can be found on figure 6 on 3 seconds extract of a spectrogram of a car passage (fig. **??**). In the case where $w_t = 1$ second , 3 elements are therefore extracted of the spectrogram while in the case where $w_t = all$, all the spectrogram is reduced to one element (fig. **??**).

Each k element of $\mathbf{W}$ is normalized such as $||\mathbf{W_k}|| = 1$ with $|| \bullet ||$ the ℓ-1 norm.
Table 1 summarizes the parameters and their related values.

Parameter	value		
K	25	50	100
$\mathbf{w_t}$ (s)	0.5	1	*all*

Table 1: Summary of the dictionary parameters

3.2.2 Performed NMF

Supervised and semi-supervised NMF are performed for 400 iterations which is sufficiently enough to get a stabilized reconstruction. ThC-NMF is performed on a lower number of iteration (60) to prevent $\mathbf{W}$ to not deviate to much from the initial dictionary. The spectrogram $\mathbf{V}$

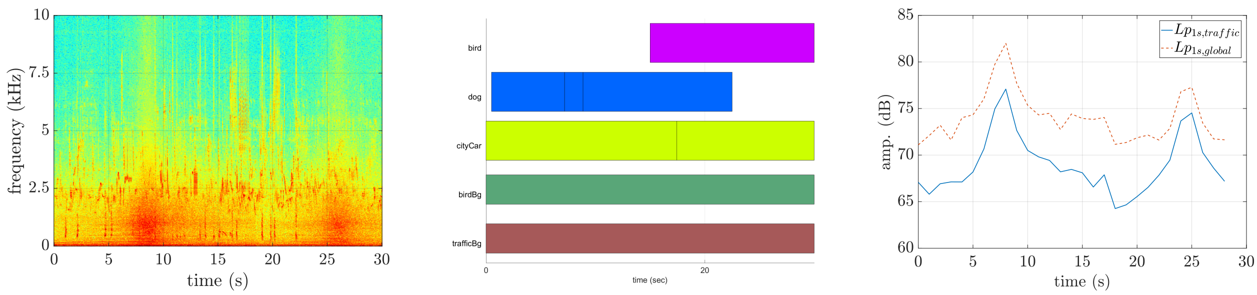
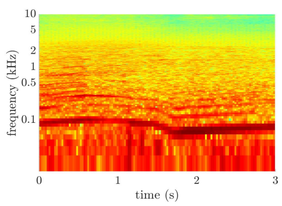
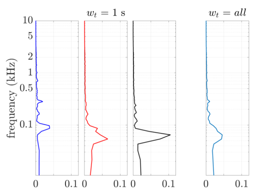


Figure 4: Example of a scene of the *animals* sub corpus. Spectrogram (on left), *Piano Roll* of the different sound classes (on the middle) and 1-s equivalent sound level of the traffic, $Lp_{1s,traffic}$ and of the global sound scene, $Lp_{1s,global}$(on right)



(a) Sample of 3 seconds of a car passage



(b) Elements of the dictionary got for $w_t = 1$ second (3 spectrum) and for $w_t = all$ (1 spectra)

and the dictionary $\mathbf{W}$ are expressed on two different formats: with a linear frequency scale ($\Delta f \approx 10.8$ Hz) and with third octave bands (29 bands). Theses two methods are considered to compare a fine grain approach (the linear scale) with a coarser 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. Furthermore, in the case of the linear frequency scale and for supervised and semi-supervised NMF, $\mathbf{V}$ and $\mathbf{W}$ are filtered at the frequencies f_c in order to focused the reconstruction of the signal of the low frequency bins. Nevertheless, 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 done with the original dictionary $\mathbf{W}$, as

$$\tilde{\mathbf{V}}_{traffic} = [\mathbf{W}\mathbf{H}_{f_c}]_{traffic}. \qquad (14)$$

With the third octave frequency scale, only the case $f_c = 20$ kHz is applied. Finally, for the ThC NMF, the threshold is define between 0.20 and 0.70 with a 0.01 step. Table 2 summarizes the parameters and the related values.

3.2.3 Metrics

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,

$$MAE = \frac{\sum_{n=1}^{N} |L^n_{p,traffic} - \tilde{L}^n_{p,traffic}|}{N}. \qquad (15)$$

- 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,

$$nRMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{p^t_{1s,traffic} - \tilde{p}^t_{1s,traffic}}{p^t_{1s,global}}\right)^2} \qquad (16)$$

where 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, 14040 settings are performed between the different form of the dictionary $\mathbf{W}$ (table 1) and the multiple parameters taken into account by NMF.

Parameter	value					
TIR (dB)	-12	-6	0	6	12	
sub-classes	alert	animals	climate	humans	transportation	mechanics
spectral representation	linear			third octave		
β	1			2		
$\mathbf{f_c}$ (kHz)	0.5	1	2	5	10	20
method	filter		supervised NMF	semi-supervised NMF	ThC NMF	
t	0.20:0.01:0.70					

Table 2: Summary of the different parameters taken into account in the frequency low-pass filter and NMF process and their values for the estimation of the traffic sound level

4 Results

Table 3 summarized, according to the 3 main parameters (method, β and spectral representation), the MAE error averaged on all sub-classes and all TIR.

The two first lines resume the error produced by the filter. $f_c = 20$ kHz is equivalent to consider all the sound mixtures without distinguish the traffic from the others sound sources. Consequently , in low TIR (-12 and -6), where traffic component is discreet, the error is more important than in high TIR (6 and 12) where the traffic component is predominant. $f_c = 500$ Hz is the cut-off frequency with the lower mean error. It is then the baseline to refer to compare the performance of NMF. In low TIR, for *alert* and *animals*, sub-classes composed of higher frequencies, the impact of the filter is quite interesting as it suppress theses frequency components whereas for the other sub-classes where low frequencies are presents, the error as higher as the filter cannot dissociate the traffic element from the other sound sources. The sounds levels are then overestimate (figure 6a). In opposite, in high TIR, the low-pass filter removes too much energy from the traffic which has the consequence to underestimate the sound levels (figure 6b).

The results of different versions of NMF are summarized next to the filter results with, in bold, the NMF errors lower than the baseline. With the exception of supervised NMF with third octave spectral representation, all NMF have a lower mean error than the baseline. However, the best combination is got with a ThC NMF with a third octave spectral representation, $\beta = 2$ and a threshold $t = 0.54$. For each method (supervised, semi-supervised and ThC NMF), the best scenario is detailed according the sub-classes and the TIR (Figure 7a, 7b, 7c and 2).

In the case of supervised NMF, on all the TIR, if the mean score is lower than the baseline method, it offers similar performances as for $TIR = [0, 6, 12]$, the mean score are nearly close. The performance of supervised NMF is mainly visible for $TIR = -12$ where the error with the baseline decrease by more than 1 dB.

The gain of the semi-supervised approach is most of all significant for these low TIR where the add of a mobile part, $\mathbf{W_r}$, into $\mathbf{W}$ allows to take into account the other predominant sound sources (example in Figure 8a) while supervised NMF is constrained to use traffic elements to reduce the distance/divergence between $\mathbf{V}$ and $\mathbf{WH}$. This improvement, according to supervised NMF results, can be seen in Figure 7c where the error on each sub class for $TIR = -12$ and $TIR =$ -6 are much lower. Nevertheless, these degrees of freedom are restrictive for high TIR. Indeed, without constraint semi-supervised NMF is free to include traffic components in W_r decreasing the rebuilt of the traffic component (Figure 8b). The error are then increasing for theses TIR. In the opposite, supervised NMF has to use the *traffic* element present in W which improve the performances.

Finally, ThC NMF with a threshold fixed at $t = 0.54$ offers the most lower results. The dictionary update allows to the method to adapt to the different TIR and to only keep the closest elements of the *traffic* component 9.

The gain is significant on all TIR with the exception of $TIR = 0$. On figure **??**, we can see the improvement of . While supervised NMF have high errors on low TIR for the sub-classes *human* and *transport*, ThC NMF succeed to have error inferior to 4 dB. The error for *climate* and *mechanics* are, on the other hand, similar to those of supervised NMF.

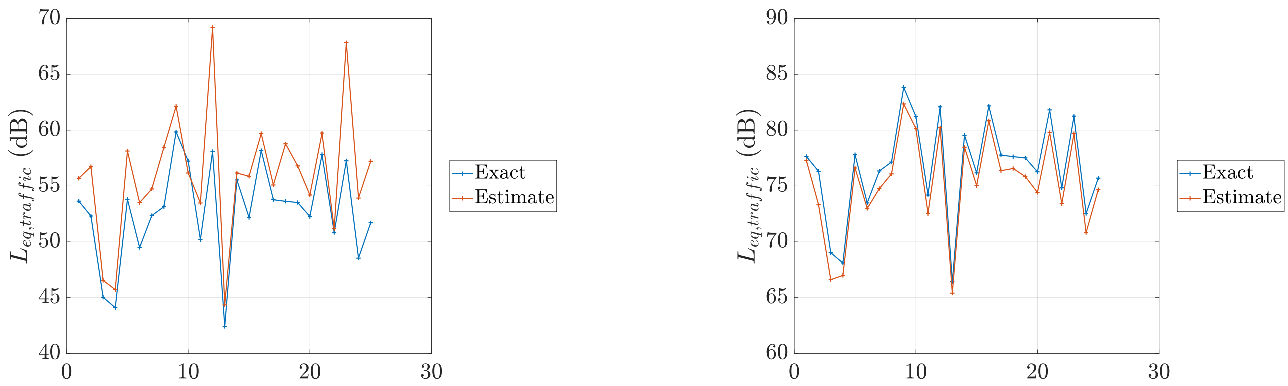
5 Conclusion

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.

K	$\mathbf{w_t}$ (s)	$\mathbf{f_c}$ (kHz)	**spectral representation**	β	**method**	**t**	MAE **(dB)**
		20			filter		4.69 ($\pm$ 4.52)
		0.5			filter		2.89 ($\pm$ 2.84)
25	all	0.5	spectra	1	supervised NMF		**2.76 ($\pm$ 2.86)**
50	0.5	0.5	spectra	2	supervised NMF		**2.56 ($\pm$ 2.48)**
50	0.5	20	third octave	1	supervised NMF		3.44 ($\pm$ 3.70)
50	0.5	20	third octave	2	supervised NMF		3.02 ($\pm$ 3.33)
100	0.5	20	spectra	1	semi-supervised NMF		**2.36 ($\pm$ 1.23)**
100	0.5	20	spectra	2	semi-supervised NMF		**2.34 ($\pm$ 1.35)**
100	0.5	20	third octave	1	semi-supervised NMF		**2.38 ($\pm$ 1.26)**
100	0.5	20	third octave	2	semi-supervised NMF		**2.43 ($\pm$ 1.43)**
			spectra	1	ThC NMF		
100	all	20	spectra	2	ThC NMF	0.37	**2.22 ($\pm$ 2.37)**
100	all	20	third octave	1	ThC NMF	0.57	**2.19 ($\pm$ 2.01)**
100	all	20	third octave	2	ThC NMF	0.54	**2.16 ($\pm$ 2.24)**

Table 3: Best results according to β, method and spectral representation



(a) (b)

Figure 6: Global sound levels of the traffic estimated by the frequency low-pass filter with $f_c = 500$ Hz for the sub-classes *alert*: at $TIR = -12$ (6a) and at TIR (6b).

[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.

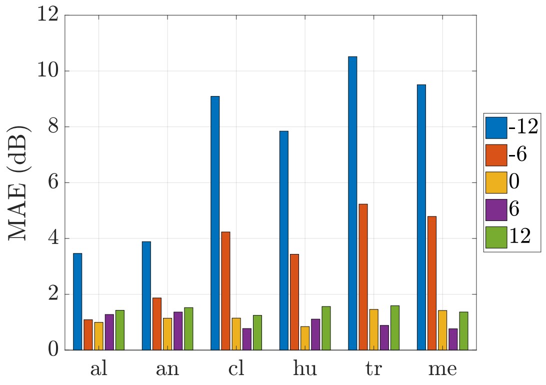
[8] C. Mietlicki, F. Mietlicki, and M. Sineau. An innovative approach for long-term environmental noise measurement: Rumeur network. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2012, pages 7119–7130. Institute of Noise Control Engineering, 2012.

[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.

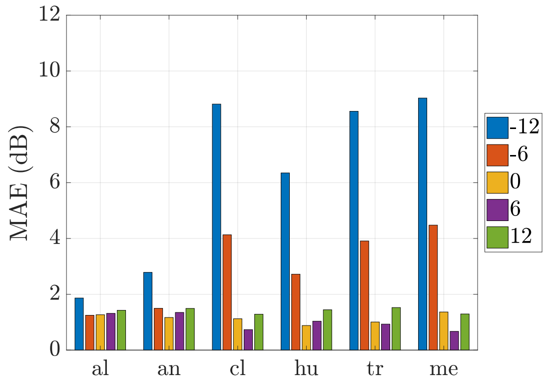
[10] D Manvell, L Ballarin Marcos, H Stapelfeldt, and R Sanz. Sadmam-combining measurements and calculations to map noise in madrid. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2004, pages 1998–2005. Institute of Noise Control Engineering, 2004.

method	filter	filter	supervised NMF	semi-supervised NMF	ThC NMF
β			2	2	2
spectral representation			spectra	spectra	third octave
f_c (kHz)	20	0.5	0.5	20	20
-12	12.25 ($\pm$ 0.05)	7.36 ($\pm$ 3.00)	**6.23 ($\pm$ 3.19)**		**5.11 ($\pm$ 3.10)**
-6	6.96 ($\pm$ 0.05)	3.44 ($\pm$ 1.65)	**3.00 ($\pm$ 1.39)**		**2.87 ($\pm$ 1.55)**
0	3.00 ($\pm$ 0.03)	1.17 ($\pm$ 0.24)	**1.14 ($\pm$ 0.17)**		1.38 ($\pm$ 0.38)
6	0.97 ($\pm$ 0.01)	1.03 ($\pm$ 0.26)	**1.01 ($\pm$ 0.28)**		**0.70 ($\pm$ 0.32)**
12	0.26 ($\pm$ 0.00)	1.45 ($\pm$ 0.13)	**1.41 ($\pm$ 0.10)**		**0.76 ($\pm$ 0.19)**

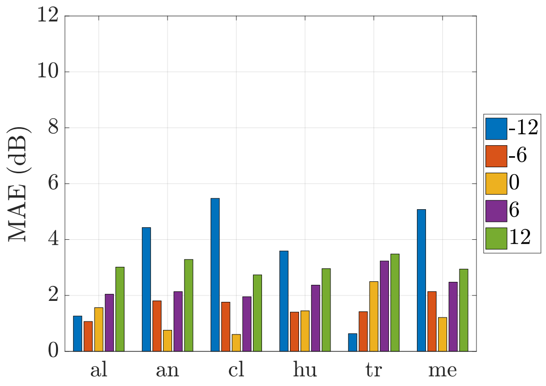
Table 4: MAE error averaged on all sub-classes on each TIR for the best scenario according to each method



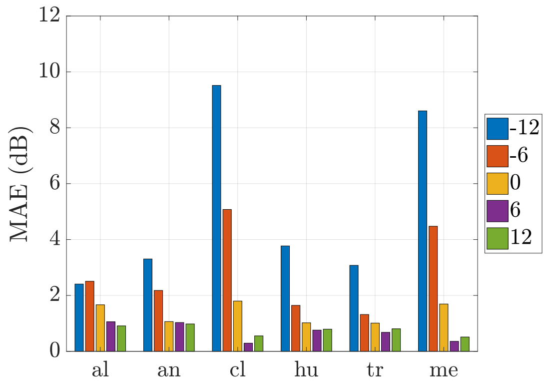
(a) filter method with $f_c = 500$ Hz



(b) Supervised NMF, $\beta = 2$, linear spectral representation, $f_c = 0.5$ kHz



(c) Semi-supervised NMF, $\beta = 2$, linear spectral representation, $f_c = 20$ kHz



(d) ThC NMF, $\beta = 2$, third octave spectral representation, $f_c = 20$ kHz

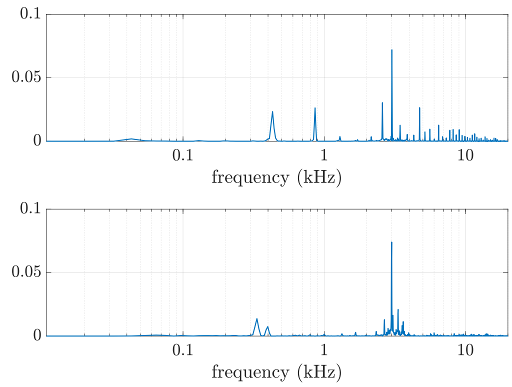
Figure 7: MAE error for each sub-class and TIR according to the the best results with the filter and each method (supervised, semi-supervised and ThC

[11] 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.

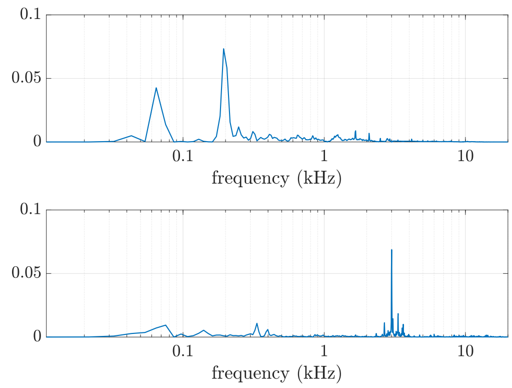
[12] L. Bellucci, P.and Peruzzi and G. Zambon. LIFE DYNAMAP project: The case study of Rome. *Applied Acoustics*, Part B(117):193–206, 2017.

[13] 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.

[14] B. Defreville, F. Pachet, C. Rosin, and P. Roy. Automatic Recognition of Urban Sound Sources. Audio Engineering So-



(a) $\mathbf{W_r}$ for an *alert* scene at TIR = -12



(b) $\mathbf{W_r}$ for an *alert* scene at TIR = 12

ciety, 2006.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

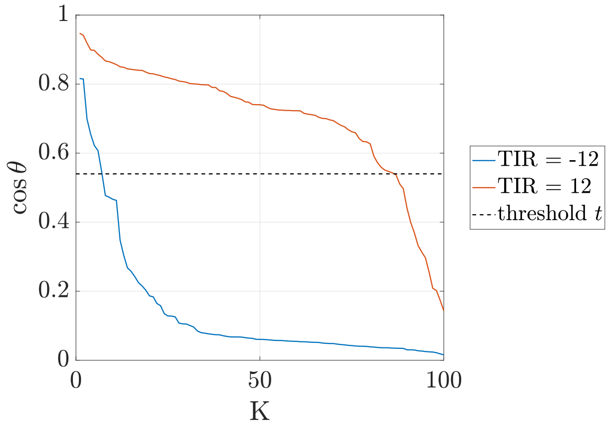


Figure 9: Example of the similarity $\cos\theta$ for an sound mixture of the *alert* sub-class for two TIR and with the threshold t: $TIR = -12$ (a) and $TIR = 12$ (b)

[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] 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.

[31] 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.

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

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

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

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

[38] M. Fornasier and H. Rauhut. Iterative thresholding algorithms. *Applied and Computational Harmonic Analysis*, 25(2):187–208, 2008.

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

[40] 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.

[41] 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.