26th International Congress on Sound and Vibration
7-11 July 2019, Montreal

STUDY OF THE NON-NEGATIVE MATRIX FACTORIZATION BEHAVIOR TO ESTIMATE THE URBAN TRAFFIC SOUND LEVELS

Jean-Rémy Gloaguen, *UMRAE, CEREMA, France, email: jean-remy.gloaguen@orange.fr*
Mathieu Lagrange, *LS2N, Ecole Centrale de Nantes, France*
Arnaud Can, *UMRAE, Ifsttar, France*
and Jean-François Petiot, *LS2N, Ecole Centrale de Nantes, France*

The advent of low-cost acoustic sensor networks in cities raises new interesting approaches for improving the monitoring of the acoustic quality of cities. Many innovative approaches are developed to improve knowledge on sound environments: sound environment recognition, sound source detection, etc. In order to improve the road traffic noise mapping, the use of a specific version of the Non-negative Matrix Factorization (NMF), named thresholded initialized NMF, as a source separation method to estimate the sound level of road traffic from measurements, has proved to be a successful approach. This paper proposes to further detail the functioning of the thresholded initialized NMF on a corpus composed of urban sound scenes mixing traffic and specific interfering components with calibrated sound levels in order to better understand its behavior according to the different sources encountered. The study reveals the different performances of this approach depending on the noise levels of the interfering sources and their proximity to the urban traffic spectrum.

Keywords: road traffic, non-negative matrix factorization, traffic noise mapping

1. Introduction

Low-cost acoustic sensor networks are currently deployed in cities to assess the urban sound environment as the DYNAMAP project [1] in Italy or the CENSE project [2] in France. The use of such networks allows innovative approaches to better estimate the soundscape through, for instance, the urban sound environment classification [3]. One application of interest is the improvement of the traffic noise mapping ordered by the European Directive 2002/EC/49. These maps are currently generated from predictive models and collected traffic data to estimate the A-weighting traffic sound level through the day, L_{DEN}, and the night, L_N. The use of these networks could facilitate the updating of the traffic maps or even the generation of dynamic maps by the assimilation of the measured data with the predicted sound levels [4]. Prior to data assimilation, the issue is to correctly estimate the traffic sound level from acoustic measurements [5, 6]. As the urban sound environment is a complex environment gathering lots of different sounds (car passages, voices, whistling bird, car horn, airplanes, etc.) that overlap in time, the estimation of the traffic sound level based on acoustic measurements is not a trivial task [7]. Although near major roads, traffic is predominant, there are many places where it overlaps with other sound sources that contribute significantly to the overall sound levels. To circumvent this issue, Socoró et al. propose an anomalous noise events detector [6]. It consists in detecting in each time frame 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 has been proposed in Gloaguen et al. [8], which is based on the blind source separation paradigm to reliably estimate the traffic noise level even when the traffic is not dominant in the mixture. It consists in separating the contribution of the traffic from the other sources within a polyphonic scene with the help of the Non-negative Matrix Factorization framework (NMF). In their study, the authors present a new approach of this method, named thresholded initialized NMF (TI NMF) whose estimate of the traffic sound level exceed in precision those of state-of-the-art approaches (supervised and semi-supervised). TI NMF then makes it possible to estimate the noise level with an estimation error less than 2 dB[1]. To do so, their study is based on simulated sound mixtures covering different sound environments (park, quiet street, noisy street, very noisy street) whose realism has been tested on a perceptual test to ensure that the sound mixtures are perceptively equivalent to audio recorded in the streets [9]. One major advantage of this approach is its application in a wide range of urban areas, even where the traffic noise is relatively low compared to the remaining contributions. In addition, NMF is well suited for monaural sensor networks. Here, to better understand the TI NMF behavior, this method is tested on a new corpus of sound mixtures mixing a traffic component with a calibrated sound level and specific sound classes as perturbators. The use of simulated sound scenes allows the authors to design a rigorous experimental validation protocol which offers a high level of control on the design of the scenes and the knowledge of the exact contribution of the traffic component ($L_{p,traffic}$). The remaining of the paper is organized as follows. Section 2 details the technical aspects of TI NMF. Section 3 describes the urban sound mixtures 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, [10], 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 $\mathbf{V} \approx \tilde{\mathbf{V}} = \mathbf{WH}$ 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 $\mathbf{W}$, thus inducing a part-based representation. The choice of the dimensions is often made so that $F \times K + K \times N < F \times N$ [11]. 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 minimized

$$\min_{\mathbf{H}\geq 0, \mathbf{W}\geq 0} D\left(\mathbf{V}||\tilde{\mathbf{V}}\right) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_\beta\left(\mathbf{V}_{fn}|\left[\mathbf{WH}\right]_{fn}\right). \tag{1}$$

The operator $d_\beta(x|y)$ is a divergence calculation and usually belongs to the $\beta-$divergence class [11] in which the well known Euclidean distance and the Kullback-Leibler divergence.

[1] Tool available at: `https://bitbucket.org/jean-remy_g/trafficsoundlevelestimation`

2.2 Thresholded initialized NMF

Thresholded initialized NMF has been proposed in [8] and is based on the unsupervised NMF. Usually in unsupervised learning, $\mathbf{W}$, as $\mathbf{H}$, is initialized randomly. In TI NMF, an initial dictionary, $\mathbf{W_0}$ dedicated to a specific sound source, is built from a learning database, see part 3.1. Then NMF is performed where $\mathbf{W}$ and $\mathbf{H}$ are updated alternatively. $\mathbf{W}$ is therefore updated with a forced initialization with *a priori* knowledge but is adapted to the actual content of the scene under study. Among the different algorithms proposed to solve 1, the multiplicative update of $\mathbf{W}$ and $\mathbf{H}$ is chosen as it has been well studied in the literature and ensures convergence of the results. It can be found in [12]. After N iterations, a measure of similarity $D\left(\mathbf{W_0}||\mathbf{W}^{'}\right)$ between $\mathbf{W_0}$ and the obtained dictionary $\mathbf{W}^{'}$ for each element k is computed using a cosine similarity metric,

$$D\left(\mathbf{W_0}_k||\mathbf{W}^{'}_k\right) = \frac{\mathbf{W_0}_k.\mathbf{W}^{'}_k}{||\mathbf{W_0}_k||.||\mathbf{W}^{'}_k||}. \tag{2}$$

$D\left(\mathbf{W_0}_k||\mathbf{W}^{'}_k\right) = 1$ means that the k-th element of $\mathbf{W}^{'}$ is identical to the k-th element of $\mathbf{W_0}$. On the contrary, $D\left(\mathbf{W_0}_k||\mathbf{W}^{'}_k\right)$ = 0 means that the elements are very different. This measure has the advantage to be bounded between 1 and 0 and to be invariant with respect to scale. The elements in $\mathbf{W}^{'}$ that can belong to $\mathbf{W}_{traffic}$ are selected by a *hard thresholding* method. It is defined as $\mathbf{W}^{'}_k \in \mathbf{W}_{k,traffic}$ if $D\left(\mathbf{W_0}_k||\mathbf{W}^{'}_k\right) > t$ where t is a fixed threshold.

3. Experimental protocol

In order to study the TI NMF behavior according urban sounds sources, the experimental protocol, the databases and the experimental factors involved in the experiment are presented, each of these experimental factors having multiples modalities, see Figure 1.

3.1 Test databases

The test database is designed with the sound scene synthesizer *SimScene* [2] [13], a simulator that creates monaural sound scenes by sequencing audio samples that come from a database of isolated sounds[14]. The database[3] used is presented in [9]. 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. This database is composed of 3 sub-corpus of 25 audio files each lasting 30 seconds. Each sub-corpus is characterized by an interfering sound class which can be *animals* (an.), composed of barking dogs and whistling birds, *humans* (hu.), composed of crowd noises and voices, and *transportation* (tr.), composed of train, tramway and plane sounds. Each sound mixture is summed with a traffic component (the sum of the road traffic background noise and the sound events generated by the *passing car* class) that makes the estimation of the traffic level more difficult. In each file, the traffic component is present. 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 calibrated level according to the sound level of the interfering class, $L_{p,interfering}$ such as $TIR = L_{p,traffic} - L_{p,interfering}$ with the *Traffic Interference Ratio* $TIR \in$ {-12, -6, 0, 6, 12} dB. The range of these values is large but in the urban environments, the TIR typically lies between -6 dB and 12 dB [9]. The case TIR = -12 dB is then an extreme case to study the NMF behavior. When $TIR < 0$ dB, the traffic component is less present

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

[3] available at: `https://zenodo.org/record/1213793`

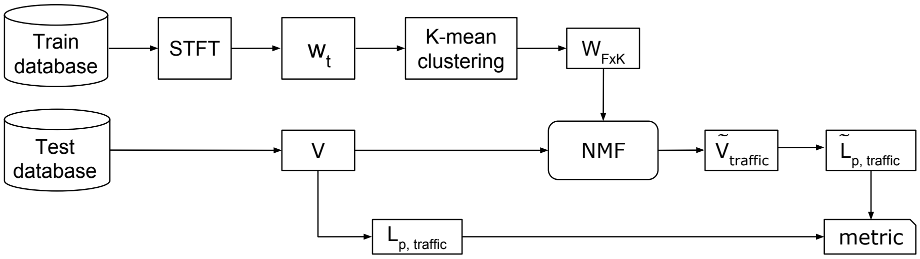


Figure 1: Block diagram of the NMF estimator with the dictionary design composed from a second sound database.

than the interfering class. On the contrary, for $TIR > 0$ dB, the traffic class is louder than the interfering class. The total number of scenes in the test database is then 375 (3 sub-corpus × 25 scenes × 5 TIR values) leading to a full duration of 3 hours and 15 minutes.

3.2 NMF experimental factors

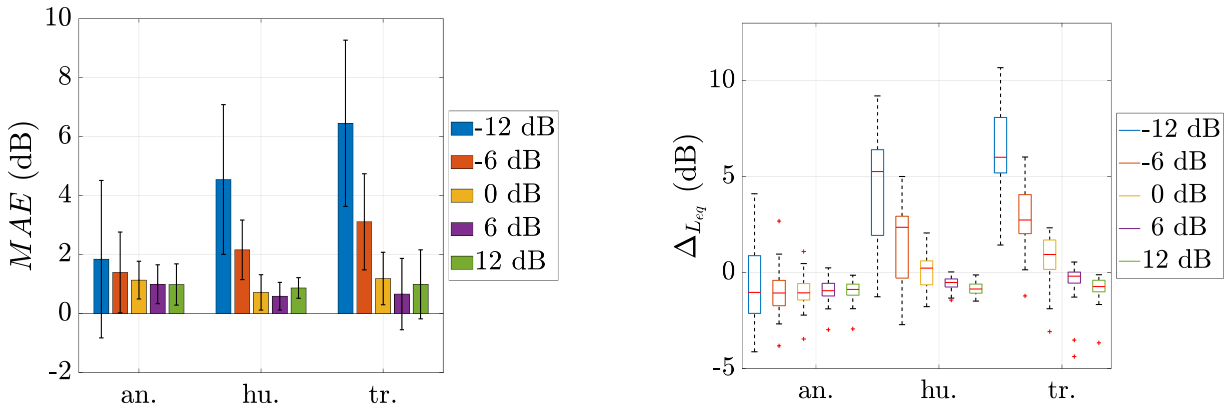
The Figure 1 presents the different steps involved in the NMF process. The dictionary design requires a train database composed of 53 audio samples of passing cars. A four-step process is realized to generate different versions of dictionaries with a sub-sampling of the audio spectrograms by a windowing of w_t seconds ($w_t \in \{0.5,\ 1\}$ second), a rms calculation of each window and a K-means clustering algorithm reducing the number of spectra to $K \in \{25,\ 50,\ 100,\ 200\}$. Each basis vector of $\mathbf{W}$ is normalized such as $||\mathbf{W_k}|| = 1$ with $|| \bullet ||$ the ℓ-1 norm. The 8 versions of the dictionary are then used as the initial dictionaries $\mathbf{W_0}$. A full description of the dictionary design can be found in [8]. 100 iterations are performed with the Euclidean distance and the Kullback-Leibler divergence. The spectrogram $\mathbf{V}$ and the dictionary $\mathbf{W}$ are expressed with third octave bands (F = 29). This coarser method allows us to reduce the dimensionality and thus decrease the computation time. Also, it is a suited representation to this sound environment as third octave bands are widely used in the urban acoustic field, compared to MFCCs for instance. The range of threshold values is set between 0 and 1 with a 0.01 increment step. Considering the experimental factors (sub-corpus, TIR, w_t, K, EUC dist./K-L div., t) derived from the different modalities of each experimental factor, 24240 settings are performed. For each setting, the estimator is performed on the 25 scenes of a sub-corpus. For one sound scene, the estimated traffic sound level, $\tilde{L}_{p,traffic}$, of the entire scene is calculated, $\tilde{L}_{p,traffic} = 20 \times \log_{10}\left(\frac{p_{rms}}{p_0}\right)$ where p_{rms} is the effective pressure deducted from the estimated traffic spectrogram $\tilde{\mathbf{V}}_{traffic}$ and p_0 is the reference sound pressure, $p_0 = 2 \times 10^{-5} Pa$. The A-weighting of the sound levels is not considered here as it decreases the low frequencies levels where the road traffic components are mainly present.

3.3 Metrics

For each setting of experimental factors, 25 values of $\tilde{L}_{p,traffic}$ are obtained and compared to the 25 exact sound levels, $L_{p,traffic}$. Its performance is assessed through the calculation of one reference metric, the Mean Absolute Error (MAE). It expresses the quality of the long-term reconstruction of the signal and is defined as

$$MAE = \frac{\sum_{m=1}^{25} |L^m_{p,traffic} - \tilde{L}^m_{p,traffic}|}{25}. \tag{3}$$

This error is calculated for each sub-corpus and each TIR value. Then a mean MAE, $mMAE$,



(a)

(b)

Figure 2: MAE errors (a) and distribution of relatives error (b) on each sub-corpus and TIR value.

is calculated on the entire test database. This metric expresses the average performance of NMF for a unique combination of modalities.

4. Results

Among all the combination of modalities, the one retained its the one generating the lowest $mMAE$ error in order to facilitate the study of the results. It is obtained with w_t = 0.5 s, K = 200, β = 2 and t = 0.42 with $mMAE$ = 2.57 (± 2.18) dB. thus, the optimal modalities found here are not exactly the same than the one found [8], as the corpuses are completely different. That being said, the TI NMF performs consistently well, motivating the authors to study further its behavior.

The errors on each sub-corpus and for each TIR value are displayed on Figure 2a. The error at TIR = -12 dB, as it is an extreme case where the traffic component is low, are the most important for each sub-corpus with high standard deviation. For *animals* and *transportation*, the MAE errors exceed 3 dB. With the exception of the *transportation* sub-corpus at $TIR = -6$ dB, the rest of the errors are inferior to 3 dB. The errors for *animals* sub-corpus are lower than the two other sub-corpuses as it contains interfering sound events relative to bird's whistles, in a highest frequencies domain than the traffic spectra, and dog barking, more brief in time. One observes that the MAE errors for this sub-corpus decrease as the TIR values increase. On the contrary, for *humans* and *transportation* sub-corpus for $TIR \leq 0$ dB, the MAE errors decrease too but increase when $TIR > 0$ dB. In parallel to the bar distribution of the MAE errors, the distribution of relative errors ($\Delta_{L_{eq}} = \tilde{L}_{eq,traffic} - L_{eq,traffic}$) is added in Figure 2b with boxplots. When $\Delta_{L_{eq}} > 0$, the traffic sound level estimated by TI NMF is over-estimated while it is underestimated when $\Delta_{\tilde{L}_{eq}} < 0$. For the *animals* sub-corpus, one notices that TI NMF underestimates the traffic sound level for all TIR values. For the two others sub-corpus, TI NMF mostly overestimates the traffic sound level when $TIR \leq 0$ dB and then underestimates it when $TIR > 0$ dB. To better understand the TI NMF behavior, the evolution of the MAE errors according to the threshold and the distance $D(\mathbf{W_0}||\mathbf{W})$ shape are displayed in Figure 3.

In Figure 3a, the distances $D(\mathbf{W_0}\|\mathbf{W})$ are sorted in descend order for each TIR values and sub-corpus. Its behavior is variable according to the sub-corpus and the TIR value. With the fixed threshold $t = 0.42$, only less than 100 elements from the obtained dictionary $\mathbf{W}$ are considered in $\mathbf{W}_{traffic}$ when TIR = -12 dB. A major part of $\mathbf{W_0}$ diverges from the original traffic spectra to simulate the interfering source. The more the traffic is predominant, the more the number of traffic elements in $\mathbf{W}_{traffic}$ growths.

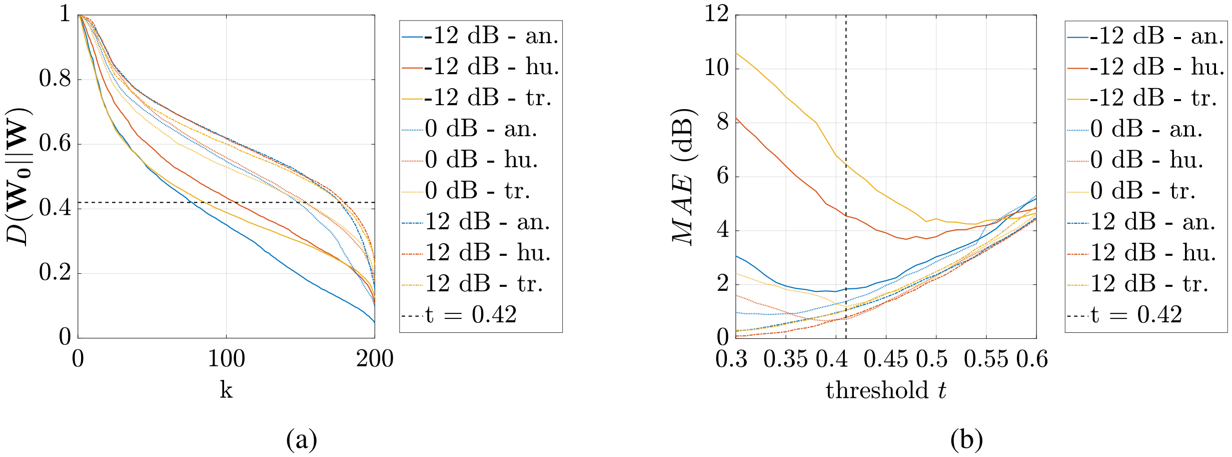


Figure 3: Distances $D(\mathbf{W_0}\|\mathbf{W})$ sorted in descend order and MAE errors evolution according threshold values.

For TIR = 12 dB, it is more than 180 elements that are considered. The traffic component is then described by TI NMF with a greater precision. The evolution of $D(\mathbf{W_0}\|\mathbf{W})$ is more influenced by the TIR values than by the interfering source as almost the same number of elements are considered in $\mathbf{W}_{traffic}$ whatever the sub-corpus. TI NMF makes it possible to adapt the dictionary by learning it for each scene. With the fixed threshold it is also a variable number of elements in the $\mathbf{W}_{traffic}$ that are considered, discarding the other elements that diverge too much from the original traffic spectra on $\mathbf{W_0}$.

Figure 3b displays the evolution of the MAE errors according to the threshold which allows us to better apprehend the estimation errors. It states the existence of optimal thresholds in each different case. The lower the TIR value, the higher the threshold has to be, while it has to decrease when the traffic component becomes predominant. At TIR = -12 dB, the optimal threshold is inferior to the fixed threshold t = 0.42 for *animals* sub-corpus while it is superior for the two other sub-corpuses. As a result, for the first case, TI NMF, with t = 0.42, does not take into account a sufficient number of elements in $\mathbf{W}_{traffic}$ which leads to an underestimation of the traffic sound level as can be seen in Figure 2b. On the opposite, as the optimal threshold for the two others classes is higher, TI NMF includes too many elements that includes some elements of the interfering sound classes. This behavior is due to the interfering sound source being more different in the *animals* sub-corpus than in the *humans* and the *transportation* sub-corpuses. Overall, this leads to an overestimation of the traffic sound level. When the TIR values increase, the traffic component becomes more important in the sound mixtures and in the dictionary $\mathbf{W}$. When TIR = 12 dB, the optimal error is obtained for a lower threshold to almost take into account all the element of the dictionary $\mathbf{W}$. TI NMF with t = 0.42 discards some traffic spectra which generates the underestimation of the traffic component.

Finally, it has to be reminded that the presence of these sound sources and the TIR values are unknown in an urban context. Despite these different behaviors and performances, TI NMF stays a good approach and makes it possible a compromise to adapt on each case.

TODO ce dernier par doit etre reecrit ou enlevé

5. Conclusion

The thresholded initialized Non-negative Matrix Factorization behavior has been studied on urban sound mixtures in order to better apprehend its performances according different urban sound sources dealing with the estimation of the traffic sound level. With this method, a dictionary $\mathbf{W_0}$ is initialized with road traffic spectra and optimized to fit the data at hand. The traffic elements that are similar to the road traffic spectra are then extracted by hard thresholding of TODO what ?. TI NMF makes it possible to learn a specific dictionary on each scene and to only consider the least divergent part of the dictionary thanks to the selection made during the thresholding step. TODO reformuler The sound database used here is an artificial build-up of a traffic component and different interfering sound classes (*animals*, *humans*, *transportation*) with calibrated sound level in order to propose a controlled framework. Among the different modalities of the experimental factors, only one combination was retained (K = 200, w_t = 0.5 s, Euclidean distance, t = 0.42). This choice was made in order to reduce the complexity of the study and to propose the most efficient method, whatever the nature of the interference sound source, so as not to add a classification step before.

With this combination, the performances are changing according to the predominance of the traffic and the different interfering sound sources. In the case where the interfering source differs from the road traffic by its spectral and temporal shapes, the errors induced on the estimation of the traffic sound level are low whatever its presence ($MAE < 2$ dB). These errors tend to decrease with the increase of the traffic presence. In this case, TI NMF tends to underestimate the traffic sound levels because of the fixed threshold that is then too high and does not allow a sufficient number of *traffic* elements to be taken into account in $\mathbf{W}_{traffic}$. In the case where the spectral shape of the sources are similar to the ones of road traffic, two types of errors appear. In the case where $TIR \leq 0$ dB, the fixed threshold is considered too low which causes the consideration of interfering elements in the traffic noise level calculation and thus its overestimation. When TIR is positive, the threshold is now considered as too high which leads to an underestimation of the traffic sound levels.

As a result, the TI NMF performances with a single fixed over the entire corpus make it possible to perform reasonably well without the need of a classification step. It then presents itself as a compromise in its performances according to the predominance of the traffic and its similarity with other interfering sound sources. If the sound scenes are composed of the traffic source and an interfering source, the urban sound environments are the result of the mix of all these sources sometimes emitted simultaneously. Keeping a fixed threshold remains necessary in order to better determine the road traffic sound level in all the cases.

The experimental protocol and the evaluated estimators have been implemented with the Matlab software. For reproducibility purposes, the code is available online[4]. The evaluation database composed of multiple samples of urban sounds is also made available[5] for the research community with interest in detection, separation and recognition tasks of urban sound sources.

REFERENCES

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

2. Picaut, J., et al. Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling, *The Journal of the Acoustical Society of America*, **141** (5), 3808–3808, (2017).

3. Maijala, P., Shuyang, Z., Heittola, T. and Virtanen, T. Environmental noise monitoring using source classification in sensors, *Applied Acoustics*, **129**, 258–267, (2018).

[4] https://github.com/jean-remyGloaguen/ICSVnmf2019

[5] https://zenodo.org/record/1145855

4. Ventura, R., Mallet, V. and Issarny, V. Assimilation of mobile phone measurements for noise mapping of a neighborhood, *The Journal of the Acoustical Society of America*, **144** (3), 1279–1292, (2018).

5. Leiba, R., Ollivier, F., Marchal, J., Misdariis, N., Marchiano, R., et al. Large array of microphones for the automatic recognition of acoustic sources in urban environment, *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 255, pp. 2662–2670, Institute of Noise Control Engineering, (2017).

6. Socoró, J. C., Alías, F. and Alsina-Pagès, R. M. An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments, *Sensors*, **17** (10), 2323, (2017).

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

8. Gloaguen, J.-R., Can, A., Lagrange, M. and Petiot, J.-F. Road traffic sound level estimation from realistic urban sound mixtures by non-negative matrix factorization, *Applied Acoustics*, **143**, 229–238, (2019).

9. Gloaguen, J.-R., Can, A., Lagrange, M. and Petiot, J.-F. Creation of a corpus of realistic urban sound scenes with controlled acoustic properties, *173rd Meeting of the Acoustical Society of America and the 8th Forum Acusticum (Acoustics' 17)*, vol. 30, pp. 4044–4044, ASA, (2017).

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

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

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

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

14. Lagrange, M., Lafay, G., Rossignol, M., Benetos, E. and Roebel, A. An evaluation framework for event detection using a morphological model of acoustic scenes, *arXiv preprint arXiv:1502.00141*, (2015).