Applied Acoustics 143 (2019) 229–238

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Road traffic sound level estimation from realistic urban sound mixtures by Non-negative Matrix Factorization

Jean-Rémy Gloaguen [a,*], Arnaud Can [a], Mathieu Lagrange [b], Jean-François Petiot [b]

[a] Ifsttar Centre de Nantes, UMRAE, Allée des Ponts et Chaussés, 44344 Bouguenais, France
[b] LS2N, 1 rue de Noë, 44331 Nantes, France

ARTICLE INFO

Article history:
Received 9 April 2018
Received in revised form 3 August 2018
Accepted 14 August 2018
Available online 27 September 2018

Keywords:
Non-negative Matrix Factorization
Urban sound environment
Road traffic sound level estimation

ABSTRACT

Experimental acoustic sensor networks are currently tested in large cities, and appear more and more as a useful tool to enrich modeled road traffic noise maps through data assimilation techniques. One challenge is to be able to isolate from the measured sound mixtures acoustic quantities of interest such as the sound level of road traffic. This task is anything but trivial because of the multiple sound sources that overlap within urban sound mixtures.

In this paper, the Non-negative Matrix Factorization (NMF) framework is developed to estimate road traffic noise levels within urban sound scenes. To evaluate the performances of the proposed approach, a synthetic corpus of sound scenes is designed, to cover most common soundscape settings, and whom realism is validated through a perceptual test. The simulated scenes reproduce then the sensor network outputs, in which the actual occurrence and sound level of each source are known.

Several variants of NMF are tested. The proposed approach, named threshold initialized NMF, appears to be the most reliable approach, allowing road traffic noise level estimation with average errors of less than 1.3 dB over the tested corpus of sound scenes.

1. Introduction

In response to the growing demand from urban dwellers for a better environment, noise mapping has been recommended as a tool to tackle noise pollution. The enactment of the European Directive 2002/EC/49 makes such maps mandatory to cities over 100 000 inhabitants. Those maps play an important informative role, establishing the distribution of the sound levels all over the cities as well as the estimation of the number of city dwellers exposed to high sound level (>55 dB(A)) [1]. Road traffic concentrates particular attention as it is the main urban source of noise annoyance. Road traffic noise maps are typically built from data collection that consist of traffic data collected on the main roads (flow rates, mean speeds and heavy vehicle ratio) and urban geographic data (building heights and location, topology, ground surfaces ...). Follows sound emission and sound propagation computational techniques, resulting in the production of the two indicators equivalent A-weighted sound levels, L_{DEN} (*Day-Evening-Night*) and L_N (*Night*) [2]. This procedure also enables drawing up action plans to reduce the noise exposure. Despite their unanimously recognized interest, noise maps suffer from some limitations. The computing cost required to produce noise maps at the city scale calls simplifications of the numerical tools and the simulation models that both generate uncertainties [3,4]. Data collection is itself also a vector of uncertainty. Moreover, the produced aggregated indicators do not model the sound levels evolution due to the traffic variations throughout the day.

Noise measurements are thus increasingly used in addition to simulation to describe urban noise environments [5–7]. Several measurement set-ups have been proposed in the last years, including mobile measurements with high quality microphones [8,9], participative sensing through dedicated smartphone applications [10,11], or the development of fixed-sensor networks. In this latter case, the sensor networks can be based either on high-quality sensors as in [12,13], or low-cost sensors as in the DYNAMAP project [14] or the CENSE project [15]. The costs and benefits of each protocol are discussed. Mobile and participatory measures increase spatial coverage at low cost, but lack temporal representativeness. Fixed networks are very reliable for measuring sound levels temporal variations, but allow only a small spatial coverage of the network. In addition, the low-cost sensors enable a wider deployment, but at the cost of increased uncertainties, the most extreme example being smartphone applications.

* Corresponding author.
E-mail address: jean-remy.gloaguen@ifsttar.fr (J.-R. Gloaguen).

https://doi.org/10.1016/j.apacoust.2018.08.018
0003-682X/

All these measurement protocols allow the combination of measures and predictions to improve the accuracy of the produced noise maps. Traffic noise maps and measurements were compared on restrictive areas in [16,17]. Wei et al. [18] modify the acoustical parameters of the simulation thanks to noise measurements, while Mallet et al. [19] call for data assimilation techniques between models and measurements to reduce the uncertainty of the produced noise maps. However, these works make the implicit assumption that the noise measurements consist mainly of road traffic. In the aim to improve road traffic noise maps, the use of measurements has first to deal with the challenge to estimate correctly the road traffic sound level. Even if road traffic is predominant on many urban areas, urban sound environments are composed of many different overlapping sound sources (passing cars, voices, footsteps, car horn, whistling birds ...), what makes the task of estimating correctly the traffic sound level within an urban sound mixture not trivial.

Many works have dealt with the classification [20,21], the detection [22,23] or the recognition [24,25] of urban sound events. In these cases, a two-step scheme is followed where audio samples are described with a set of features (Mel Frequency Cepstral Coefficient, MPEG-7 descriptors ...) and classified with the help of a classifier (Gaussian Mixtures Models, Artificial Neural Network ...) [26,27]. The classifier is learnt from a learning database and is next applied on a test database to validate the algorithms. Dedicated to the traffic, in [28], an Anomalous Event Detection, based on MFCC features, is proposed with the specific aim to improve the traffic sound estimation. It is based on the detection of unwanted sound events in order to discard them.

An other approach, followed in this paper, is to consider the blind source separation paradigm which consists in the extraction of a specific signal inside a set of mixed signals, see Fig. 1. From the different existing methods, Non-negative Matrix Factorization (NMF) [29], appears to be a relevant method for monophonic sensor networks. Many applications can be found for musical [30,31] and speech [32,33] contents. Dedicated to sound separation with environmental sounds, Immani and Kasaï [34] used NMF in a two steps sound separation with the help of time variant gain features. Dedicated to the traffic sound separation, a first study [35] has been conducted, in which diverse NMF estimation rules are compared, namely the supervised, the semi-supervised, and the threshold initialized NMF, have been applied on a large set of simulated sound scenes. This corpus mixes 6 sound categories (*alert, animals, climate, humans, mechanics, transportation*) with a traffic component calibrated to different sound levels, according to the other sound classes (in the rest of the document, these sound classes, not related to the traffic component, are resumed as the *interfering* sound class), to obtain variable traffic predominance. The diversity of this corpus was made to assess the performances and the limits of each NMF formula. However, if this study reveals the interest of NMF for urban sound environments, the assessment of its performance on a corpus of realistic sound scenes must be carried out in order to implement it on a sensor network. Design urban sound mixtures makes it possible to access to many acoustic properties as the onset and offset time and the sound level of each sound class and especially the traffic component. The realistic aspect of such a corpus is essential to obtain sound scenes similar to recordings and to validate NMF performances. However, like all simulated process, the realism of the scenes must be perceptually verified.

In this paper, an urban sound corpus based on annotated urban recordings, and whose degree of realism is assessed through a perceptual test, is designed in order to estimate the traffic sound level with the help of the NMF framework. The different NMF approaches are described in Section 2. Next, the corpus of urban sound scenes is presented in Section 3, from the sound database built-up to its validation. The experimental protocol and the results are then presented and discussed in Sections 4 and 5.

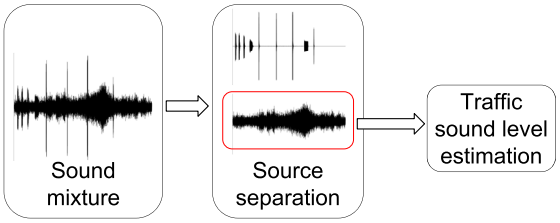

Fig. 1. Block diagram of the blind source separation model.

2. Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is a linear approximation method proposed by Paatero and Tapper [36] and popularized by Lee and Seung [29]. It consists in approximating a non negative matrix $\mathbf{V} \in \mathbf{R}^{+}_{F\times N}$ by the product of two non negative matrices: $\mathbf{W}$, called *dictionary* (or basis), and $\mathbf{H}$, called the *matrix activation* with dimensions $F \times K$ and $K \times N$ respectively,

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

The choice of the dimensions is often made such as $F \times K + K \times N < F \times N$ so that NMF can be a low rank approximation. This condition however is not mandatory. When applying NMF to audio data, $\mathbf{V}$ is usually considered as the magnitude spectrogram obtained by a Short-Time Fourier Transform, $\mathbf{W}$ includes audio spectra and $\mathbf{H}$ is equivalent to the temporal activation of each spectrum, see Fig. 2. Because of the non-negativity constraint, only additive combinations between the elements of $\mathbf{W}$ are considered.

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

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

where $D(\bullet\|\bullet)$ is a divergence calculation such as:

$$D(\mathbf{V} \,||\, \mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d_\beta\Big(\mathbf{V}_{fn} \mid [\mathbf{WH}]_{fn}\Big). \tag{3}$$

$d_\beta(x \mid y)$ is usually chosen as a β-divergence [37], a sub-classes belonging to the Bregman divergences [38] which include 3 specific divergence calculations: the Euclidean distance Eq. (4a), the Kullback–Leibler divergence Eq. (4b) and the Itakura-Saïto divergence Eq. (4c):

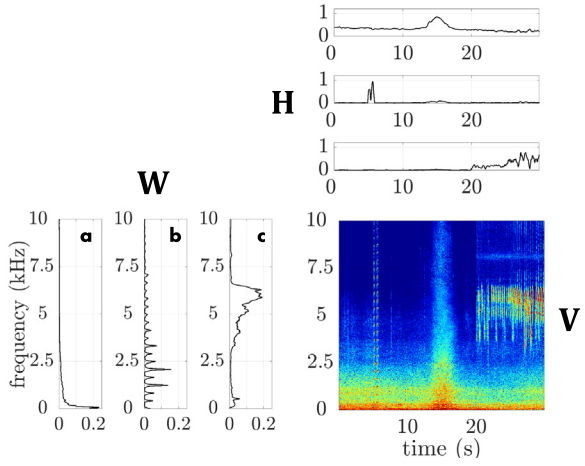

Fig. 2. NMF decomposition of an audio spectrogram V composed of 3 elements ($K = 3$): passing car (a), car horn (b) and whistling bird (c).