Reply to reviewers concerning submission JASM-D-18-00015: "Estimation of the road traffic sound levels in urban areas based on non-negative matrix factorization techniques"

September 14, 2018

As a preamble, we would like to thank the editor and the reviewer for their comments and suggestions. Following these comments, we made several changes to the article, which are summarized here. The next sections list our answers to each of the reviewers comments, with references to the revised manuscript (page, column, and paragraph) where appropriate.

1 Answers to Reviewer 1

1.1 Contribution and state of the art

1. *The authors miss some important references when they review the state of the art related to the addressed problem. For instance, those articles published within the DCASE 2017 challenge, which are not included in the paper. Moreover, other specific relevant works should be referenced, such as:*

   - *Onur Dikmen and Annamaria Mesaros. "Sound event detection using non-negative dictionaries learned from annotated overlapping events", 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), where also a threshold-based NMF approach is applied to a very similar problem.*
     And the work previously published by the same authors in DCASE 2016:
   - *Jean-Remy Gloaguen, Arnaud Can, Mathieu Lagrange, Jean-Francois Petiot. "Estimating traffic noise levels using acoustic monitoring: a preliminary study." Detection and Classification of Acoustic Scenes and Events (DCASE 2016), September 2016, Budapest, Hungary.*

   Both papers should be included as related work explicitly explaining the main differences with respect to the presented work.

→ The introduction has been enriched with the two proposed references. The differences between the present contribution and the work of Dikmen & al [2] and our previous work [4] is now clearly stated. We also added the work of Mesaros & al [7]

2. *The authors argue that "Prior to data assimiliation, the issue of the correct estimation of the traffic sound level from acoustic measurements is still unsolved [31]", being [31] a reference from 2011. The authors should rephrase the sentence or change the reference in order to update the current state of the art in this field. For instance, only reference [35] is included describing deep learning approaches, which is one of the relevant approaches in this research field during the last years.*

   → The sentence has been rephrased with updated references (please see Section 1) :'Prior to data assimilation, the issue of the correct estimation of the traffic sound level from acoustic measurements begin to be studied [6][9]. As the urban sound environment is a complex environment gathering lots of different sounds (car passages, voices, whistling bird, car horn, airplanes. . .) that overlap, the traffic sound level estimation based on measurements is not a trivial task.'

3. *As a final suggestion, I would recommend the authors completing the review of the state of the art by including a specific section about the Related Work in the manuscript. Although this section is not mandatory according to the guide for authors in this journal, it would help to clarify the contribution of the authors with respect to previous similar approaches.*

   → This section has been added as Section 2 of the updated manuscript where related work is now presented and where we highlight the differences between those works and the approach taken in our contribution.

1.2 Metric and results

1. *Since Lp values are in dB, you cannot average them linearly as considered in your proposed metric MAE of eq. (15). The values must be averaged in their linear representation before converting them into dBs again. This error may have a dramatic impact on your results.*
   →

   The use of arithmetic operators on logarithmic values is a common practice in the urban acoustics field when dealing with errors (see Aumond et al. "Kriging-based spatial interpolation from measurements for sound level mapping in urban areas." [1] and Morillas et al. "Uncertainty evaluation of continuous noise sampling." [8]). In order to better position our design choice in terms of metrics a paragraph is now added in Section 4.2.3.

   We acknowledge that strictly physically speaking this might be considered as a mistake but we chose to follow this approach as we believe it allows the calibration of the traffic sound level and to monitor metrics that are less sensitive to high values. As the errors are treated in a statistical point of view we consider that this operation is well suited.

2. *First, the results obtained when considering fc=20Khz ("fc = 20 kHz is equivalent to consider all the sound mixtures without distinction between traffic and others sound sources") show the lowest MAE for TIR=6,12, where "the traffic is the main sound source". –¿ Have the authors realized the consequences of these results? From these them someone can conclude that it is better to process nothing when interfering sources are 6 and 12 dB lower with respect to the traffic noise level!!*

   → This results suggests that the user does not apply any treatment in cases where road traffic predominates. This is not inconsistent with the purpose of this paper for two reasons: (i) a user would not call for signal processing in situations where road traffic is largely dominating such as in the vicinity of a ring road, (ii) the TIR value is an unknown parameter: the high errors of the filter for negative TIR values discredit its use. We made the precision : 'Even if the performances for the positive TIR values are high, it has to be reminded that in practice, the TIR value is not known. In consequence, without any prior knowledge, this approach cannot be applied.' (please see the beginning of Section 5).

3. *Second, for the remaining studied scenarios, i.e., TIR=-12,-6,0, the lowest averaged MAE values are obtained by the Sem-NMF, and not by the proposed TI-NMF. Therefore, from these results, it cannot be concluded that TI-NMF is the best approach in the considered experimental framework. Even more, if you remove from Table 4 the aforementioned values obtained with TIR=-12,-6, Sem-NMF becomes the best approach with an averaged MAE of 2.12, whereas TI-NMF yield an averaged MAE of 3.14, following the same analysis scheme proposed by the authors.*

   → We made a precision at the end of section 2 : 'The aim is to find the type of NMF that gives the minimal error on the reconstruction of the traffic noise signal on the whole corpus to be applied to all the different possible urban noise environments.'

   Furthermore, as now explained at the end of section 4.1, in urban sound environments, TIRs are mainly focused in the range $[-6, 12]$ dB [5]. TIR = -12 dB is a rarer case in urban environment but it is still added in order to show the limits of the evaluated methods.

4. *Finally, I would recommend the authors to consider including a statistical analysis to prove that differences between the different approaches are statistically significant or not.*

   → As suggested by the reviewers, we added a Student t-test in Section 5 with the results exposed in Table 3. This test revealed that the distribution of the traffic sound level estimations between the best methods exposed in Table 2 are statistically different.

5. *Last but not least, please review the results of the low pass filter approach with fc=20kHz in Table 4 and 3, since the global averaged MAE in Table is 4.54, while in Table 3 is 4.69.*

   → The value at TIR = 6 was a misprint, it has been corrected.

1.3 Meaningfulness of the experiments

1. *The baseline approach considered by the authors is a low-pass filter with different cut-off frequencies "fc=500, 1k, 2k, 5k, 10k, 20k Hz.". First, this baseline may be understood as too simple for the problem at hand by the research community since, at least, supervised NMF has been already considered in the literature. Second, no comparison with [38] or other previous attempts to apply NMF to the problem at hand is provided. To this aim, it would be especially interesting the comparison with (Dikmen and Mesaros, 2013). Finally, after introducing the fc sweep in section 3.2, only fc=500, 20kHz are the subsequently considered with no further justification.*

   → The baseline of the low-pass filter, even if it seems too simple, is very effective in some cases. As in the environmental acoustics community, this method would have been easily seen as a tool, we decided to consider it. We summarized all the results for the low pass filter according to the cut-off frequency in order to be more complete (Figure 9). As for the comparison with (Dikmen and Mesaros, 2013), even if they use the NMF framework, they develop a detection tool that is not adapted to this study. We believe that it was wiser to use a low-pass filter as a low complexity baseline.

2. *The authors have considered a database composed of artificially mixed sound sources. They argue this decision as follows: "The use of simulated sound scenes is mandatory for rigorous experimental validation as it offers a high level of control on the design of the scenes" –> I would suggest removing the term "mandatory" from the argumentation. Maybe the sentence could be rewritten as: "The use of simulated sound scenes allows"*

   → The sentence have been rephrased to : 'The use of simulated sound scenes allows rigorous experimental validation . . . ' (section 2).

3. *Moreover, for this reviewer it is not clear what is considered traffic and not traffic by the authors in order to compute $Lp_t raffic$. The "environmental sound scene corpus" is composed of "constant traffic noise" and "passing cars", which have been recorded in a real life environment by the authors. The car pass-bys are then used in several figures. For instance, Fig.4 shows an example created with the SimScene software showing background road traffic noise together with three overlapped sound sources: car horn, car passage and whistling birds. Does this mean that the car pass-by is*

considered as an interfering noise source? I would recommend the authors to clarify this point since a car pass-by is also traffic noise.

→ This important point is now clarified in the manuscript: 'What is called traffic component is the sum of the road traffic background noise and the sound events generated by the passing car class. On the contrary, the *interfering* sound class includes all the other sound sources not related to it. Car horn sound class belongs to this component as it is considered as a warning signal' (Section 4.1).

4. *Furthermore, the authors provide a lot of information about the configurations of the different approaches. However, I would recommend including specific information about the total duration of the train and test databases (e.g., not only saying "750 sound mixtures"), besides explicitly explaining the test scheme considered to guarantee that train and test data is decoupled when evaluating the considered sub-classes.*

   → Those numbers are now added to the manuscript. For the test database, see section 4.2.1: 'each scene during 30 seconds, it leads to a full duration of 6 hours and 30 minutes.' and for the train database, see section 4.1: 'The train database is composed of 53 audio files of isolated passing cars with a 18 minutes cumulative duration.'

5. *Finally, it is worth mentioning that considering TIR=-6,-12,0,6,12 may yield to conclusions far from what can be found in real-life environments, as discussed in the literature (e.g. see [42]). Thus, I would recommend rephrasing the following sentence: "This database allows the creation of a wide diversity of realistic urban sound scenes from the road traffic point of view [16]" accordingly*

   → The sentence is now rephrased in section 4.1 to: 'This database allows the creation of a wide diversity of urban sound scenes from the road traffic point of view' ... In addition, at the end of section 4.1 the motivation behind the choice of the range of TIR values is now discussed in depth, highlighting the fact that, indeed, the range is definitely wider than what can be found in realistic settings. The results are also now discussed by considering this matter into account.

1.4 Minors comments

1. *In the introduction, the authors explain that road traffic noise maps show Lden and LN considering the A-weighted equivalent sound pressure level, thus, expressing the noise levels in dBA. However, their work only focuses on the equivalent sound pressure level in dB, why?*

   → This choice is now explained in section 4.2.2: '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.'

2. *Sections 2.1, 2.2 and 2.3 describing NMF, Supervised and Semi-Supervised NMF could be removed after extending the related work. This way, the manuscript could emphasize the proposal TI-NMF with respect to similar works.*

   → Our contribution is targeted in part to the urban acoustic community. As NMF is not a well-known tool in this community, we thus prefer to keep the presentation of the fundamentals of this method.

3. *The authors state that "Sup-NMF and Sem-NMF updates are computed for 400 iterations, which is sufficient to reach convergence. TI-NMF is performed on a lower number of iterations (60) to prevent W to not deviate too much from the initial dictionary." –¿ Please, justify this qualitative argumentation.*

   → We now consider 100 as the number of iterations for each method, see Section 4.2.2: '100 iterations are performed for every NMF types'.

4. *Different kind of space reduction and data representation are included to reduce the computational time, e.g. K-means and third octave bands. –¿ I would recommend the authors to include some kind of diagram to better explain their approach, and, if possible, to include some experiments to argue their argumentation, for instance, to explain why you have chosen third octave bands in front of other possible parameterization such as MFCC.*

   → About the choice of third-octave bands, we specify in section 4.2.2 'The spectrogram **V** and the dictionary **W** are expressed with third octave bands ($F = 29$). This coarser method allows us to reduce the dimensionality and then decrease the computation time. Furthermore, by expressing the frequency axis on a log frequency axis, the low frequencies, where the traffic energy is focused, are described more finely than the high frequencies. Experimental validation consistently showed that considering third octave bands do not impact the performance of the estimator studied in this paper. But, most of all, it is a suited representation to this sound environment as this kind of representation is widely used in the urban acoustic field, compare to MFCC for instance.'. Also, we add a block diagram which details the dictionary building and makes clearer these different steps (please see Figure 7).

5. *"The threshold t is set between 0.30 and 0.70 with a step of 0.01" –¿ Please include some argumentation or reference to justify the range selected for the experiments beyond the ones reported in Table 3. Please, include more information of these experiments.*

   → This matter is now discussed in Section 4.2.2. 'For TI-NMF, a preliminary study showed that the range of threshold values can be evaluated between 0.30 and 0.70. An increment step of 0.01 has been considered as being sufficiently precise.'

6. *All the analyses are done considering t=0.42 as the best threshold for the TI-NMF approach. Nevertheless, the authors stand that "With a low threshold, it is possible to decrease the error", and present some examples subsequently to demonstrate that for high TIR=12 dB, the averaged MAE can be decreased by reducing the threshold value from t=0.42 to t=0.30. Correspondingly, the threshold should be moved from t=0.42 to t=0.55 to improve the results for low TIR=-12 dB. To what extent these values show better performance or they are due to overfitting?*

   → This improvement is due to a decrease of the number of elements considered as traffic components. As in TIR = -12 dB, the interfering class is more present, it takes more basis in W'. By increasing the threshold, we put aside more elements from W' that can belong to this sound class. It is equivalent, in detection task, to a reduction of the number of false-positive. It is now discussed further in the end of the section 5: 'It would be possible to decrease the error by adapting the threshold according to the TIR. For instance at TIR = 12 dB, the average error on all the sub-classes would decrease to 0.30 (± 0.11) with $t = 0.34$ where 191 elements are considered as traffic component. In the contrary, at TIR = -12 dB, by increasing the threshold to 0.53, the error would decrease to 4.49 (± 1.75) with in average 73 elements considered as *traffic* elements.'

7. *I would recommend including all the tuning and sweep analysis within an Appendix at the end of the manuscript.*

   → As lots of results have been generated by those experiments, we chose to detail only some of the parameters that may influence the performance of the studied algorithms.

1.5 Some further details

1. *A -12 is missing in the enumeration of TIR values after eq. (12). Nevertheless, I would recommend removing this enumeration from the explanation of the equation since these values should be included in the Experiments section, where they are indeed included.*

   → This part has been removed.

2. *Tables 1 and 2 could be fused within the same Table.*

   → The Tables have been fused.

3. *In Section 3.2.2, wt 0.5 1 –¿ a comma is missing between 0.5 and 1, i.e., 0.5, 1*

   → The misprint is now corrected.

4. *Fig 8. The y-axis range does not show the minimum values of the alarm Lp,1s lower than 50 dB. Please, include them to better understand the impact of the alarm on the Lp computation. Moreover, to better understand*

the example, I would recommend the authors to append the spectrogram of the example to the Lp.1s (dB) figure.

→ Outside these emergences, the car horn sound level is null, that is why we limited the y-axis range to [45 85] dB to see, at best, the curves. In addition, we followed the reviewer's recommendation by adding the spectrogram of the audio scene where the car horn sounding is visible (please see Figure 11).

5. *Fig 9. Caption: "according to the the best" –> remove one 'the*

   → The caption title has been fixed.

1.6 English writing

1. *Section 1: "Prior to data assimiliation," –> assimilation*

   → The typo is corrected.

2. *Section 2.4: "Usually in unsupervised," a name is missing in the sentence.*

   → The sentence is now: 'Usually in unsupervised learning,'

3. *Section 4: "First, the errors produced by the filter are detailed" –¿ the errors are not produce by the filter itself. Please, rephrase the sentence to better explain the readers what are referring to.*

   → The sentence is now rephrased to: 'First, the errors calculated from the traffic sound estimation of the low-pass filter are detailed'

4. *"the traffic elements selected in TI-NMF do not activated when" –¿ are not activated?*

   → The sentence is now rephrased to: 'the traffic elements selected in TI-NMF are not activated'

5. *Section 5. "In this work the non negative matrix factorization framework was used" –¿ I would recommend using present perfect instead of past tense in the conclusions.*

   → The sentence is now rephrased to: 'In this work the non negative matrix factorization framework is used . . . '

2 Answers to Reviewer 2

1. *As the previous work has been already published, the authors should provide references to it and clearly state the new contributions of the presented paper. I believe there may be a substantial overlap with the authors' paper Gloaguena, J. R., Cana, A., Lagrangeb, M., & Petiotb, J. F. Estimation du niveau sonore du trafic routier au sein de mixtures sonores urbaines par la Factorisation en Matrices Non négatives, 2018.*

→ The paper 'Gloaguen, J. R., Can, A., Lagrange, M., & Petiot, J. F. Estimation du niveau sonore du trafic routier au sein de mixtures sonores urbaines par la Factorisation en Matrices Non négatives, 2018' is a congress article without review submitted after the submission of the present paper. It deals with a similar approach but on a different sound corpus. Furthermore, as it is written in french we chose to alternatively cite 'Gloaguen, J.R., Can, A., Lagrange, M., Petiot, J.F.: Estimating traffic noise levels using acoustic monitoring: A preliminary study' which is written in English and deals with the same corpus.

2. *Some sentences in the abstract are not well formulated and are hard to follow. E.g. "The task being to the best of our knowledge never been considered in the literature, we propose an experimental protocol to validate the studied approaches that complies with standard reproducible research recommendations." should be splitted, "being" should be omitted; "raise this kind of innovative approaches" - the verb "raise" is confusing; "based on non-negative factorization framework" - the word "matrix" is missing. The authors may want to be more specific about the obtained results.*

   → The abstract have been rewritten to take those requests into account: 'The task is to the best of our knowledge never been considered in the literature. We propose...', 'Among the technical issues that bring this kind of innovative approaches,...', 'several techniques based on the non-negative matrix factorization framework'.

3. *"Their spectra can be obtained and be a basis of W." The procedure is not clear and is only described in the experimental part. This sentence seems to mean that the matrix W is comprised of exemplars (e.g. see $http://www.cs.tut.fi/sgn/arg/music/tuomasv/chime-enhancement/pP4_gemmeke.pdf$), which is not the case. In supervised NMF framework, the basis matrix W is usually pre-trained applying NMF to the training data. It would be interesting to compare such approach to the proposed K-means procedure.*

   → A block diagram has been added to better explain the dictionary learning process shown in Figure 7 of the updated manuscript. Performing an NMF training have also been tested with no significant improvements.

4. *One type of vehicles (passenger car) is addressed in the simulations. Authors may want to comment on why other types like motorcycles, which generate a palpable amount of noise and may have different acoustic characteristics, are not considered.*

   → This important limitation of the presented study is now clearly stated as future work of interest in the conclusion : 'It has to be noticed that the method would gain to be tested on larger dataset with, for instance, motorcycles and other noisy traffic elements which are not considered in this study.'

5. *Is also interesting to know if the distribution of microphones in a realistic noise level measurement scenario is taken into account during the simulations.*

   → The purpose of this study is to propose a tool that could be integrated into a sensor network. The distribution of the microphone is not a question dealt with in this work. The proposed method is designed to perform with a monophonic input. That being said, considering the inputs of multiple sensors may help to achieve more consistent estimates. This is left for potential future work.

6. *"The obtained dictionary is expressed with third octave bands", "The spectrogram V, as the dictionary W, is expressed with third octave bands (F = 29)." - please, clarify. Perhaps is better to join the description of this representation with the section 4.2.1.*

   → The manuscript have been updated, see Section 4.2.2: '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 then decrease the computation time. Furthermore, by expressing the frequency axis on a log frequency axis, the low frequencies, where the traffic energy is focused, are described more finely than the high frequencies. Experimental validation consistently showed that considering third octave bands do not impact the performance of the estimator studied in this paper. But, most of all, it is a suited representation to this sound environment as this kind of representation is widely used in the urban acoustic field, compared to MFCCs for instance.'

7. *As for the sound level calculation itself, it is not clear if A-weighting is applied.*

   → As now explicitly stated in Section 4.2.2, '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.'

8. *Also, the terminology is a bit confusing. E.g. Figure 8 "1 second equivalent sound pressure level" - to my understanding, the equivalent sound pressure level Leq is an average over the specified period of time and, therefore, would correspond to a single value. If the goal is to eventually simulate the output of a sound pressure level meter, some calibration would be needed.*

   → The caption of the Figure 8 has been corrected to: 'Evolution of 1 second equivalent sound pressure levels'. The calibration step is here not necessary as all the sound levels are set relatively.

9. *It is not clear what is the number of iterations that was used for each NMF version. It might make sense to use the reduced number of iterations for Sup-NMF as it is done for TI-NMF.*

   → To better compare the different NMF version, 100 iterations are now performed for every NMF types, see section 4.2.2.

10. *Is it true that the dictionaries W in Sup-NMF, Ws in Sem-NMF and W0 in TI-NMF are the same? Authors may want to state it clearly in Section 4.2.1.*

    → Yes it is, it has been clarified in the updated version of the manuscript in Section 4.2.1: 'The 10 versions of the built dictionary are then used for NMF. In Sup-NMF, these 10 versions correspond to $\mathbf{W}$, in Sem-NMF, they correspond to the fixed part $\mathbf{W_s}$ and for TI-NMF, they are the initial dictionaries $\mathbf{W_0}$ that are next updated.'

2.1 Minor points

1. *”at fixed stations spread all over the cities [31] [30], which would make available of the long-term evolution of the traffic noise” - better ”which would make the long-term evolution of the traffic noise available”.*

   → The sentence is rewritten.

2. *”A two-step process is generally followed :” - there is an extra space before the word ”followed”*

   → Fixed.

3. *”In the opposite” - better ”On the contrary”*

   → Fixed.

4. *”to some extend” should be ”to some extent*

   → Fixed.

5. *”by forcing its initiation” should be ”by forcing its initialization”*

   → Fixed.

6. *Figure 2 caption, please change to ”of an urban”, ”correspond to”.*

   → Fixed.

7. *Figure 3 caption, the sentence ”The 82-nd first elements are considered as traffic component.” is confusing.*

   → The sentence is rewritten.

8. *Figure 5 caption, ”experience” should be ”experiment”?*

   → Indeed, fixed.

9. *”The choice of the dimensions is often made so that $F \;\; K + K \;\; N < F \;\; N$.” - could you please provide a reference for this statement?*

   → The reference [3] is added.

10. *”must of all” should probably be ”most of all”*

    → Indeed, fixed.

11. *"The errors are too important for low TIR and this for all the sub-classes." - the sentence is not clear.*

    → The sentence is rewritten to: 'Errors are too important for all sub-classes at low TIR.'

References

[1] Aumond, P., Can, A., Mallet, V., De Coensel, B., Ribeiro, C., Botteldooren, D., Lavandier, C.: Kriging-based spatial interpolation from measurements for sound level mapping in urban areas. The Journal of the Acoustical Society of America **143**(5), 2847–2857 (2018)

[2] Dikmen, O., Mesaros, A.: Sound event detection using non-negative dictionaries learned from annotated overlapping events. In: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on, pp. 1–4. IEEE (2013)

[3] Févotte, C., Bertin, N., Durrieu, J.: Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Computation **21**(3), 793–830 (2009). DOI 10.1162/neco.2008.04-08-771

[4] Gloaguen, J.R., Can, A., Lagrange, M., Petiot, J.F.: Estimating traffic noise levels using acoustic monitoring: A preliminary study. In: DCASE 2016, Detection and Classification of Acoustic Scenes and Events, p. 4p (2016)

[5] Gloaguen, J.R., Can, A., Lagrange, M., Petiot, J.F.: 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 (2017). DOI 10.1121/1.4989346

[6] 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. In: INTER-NOISE and NOISE-CON Congress and Conference Proceedings, vol. 255, pp. 2662–2670. Institute of Noise Control Engineering (2017)

[7] Mesaros, A., Heittola, T., Dikmen, O., Virtanen, T.: 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), pp. 151–155 (2015). DOI 10.1109/ICASSP.2015.7177950

[8] Morillas, J.B., Gajardo, C.P.: Uncertainty evaluation of continuous noise sampling. Applied Acoustics **75**, 27–36 (2014)

[9] Socor, J.C., Alas, F., Alsina-Pags, 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). DOI 10.3390/s17102323