Reply to reviewers concerning submission AAA-D-19-00080: "Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques"

October 7, 2019

As a preamble, we would like to thank the editor and the reviewers for their comments and suggestions. Following these comments, we made several changes to the article including a modification of the structure of the document, 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 I

1. *My biggest concern with this work is its generalization. By training on a dataset derived from only 19 recordings, and evaluating on only 6 recordings that weren't replicated in the synthetic data, it's unconvincing that this model would generalize to new data.*

   → Statistics used to simulate sound scenes are based on 74 recordings of which the 19 used for replication are a subset. Section 2.1 has been rewritten to describe more extensively this reference database. To generate more diverse scenarios the variance of all statistics is next further increased. However, we acknowledge that the model needs to be evaluated on a more complete and representative set of recordings to assess fully its generalization capabilities.

2. *It's even less convincing considering that in the final model, only the presence of birds was shown to be predictive of pleasantness, when in large cities like NYC, construction, loud music, dogs, and overnight delivery trucks are some of the most complained about sources of noise, none of which are even included in this model. Therefore, while I don't believe that the authors approaches and analyses are flawed, I think the datasets they are basing this research on are very limited and not representative of urban*

acoustics. [...]
Section 4.2 - Seems pretty odd that only birds were used in the model, no?

→ In this paper, we chose to focus on the three main canonical categories of sources that are known to impact soundscape quality. When applying the proposed approach to a specific environment, the taxonomy of sound sources needs to be extended accordingly. This is now discussed in Section 5. Also, as now emphasized in the Introduction and Section 2, the present study considers the point of view of pedestrians, which is vastly different from that of residents in terms of sources of interest.
The quality of the pleasantness model obtained from time of presence approximations is bound by the listening test corpus, and thus mostly by the contents of simulated scenes. Section 4.1 discusses possible causes to the absence of other sources in the model. However, this model is not intended as a contribution but rather as a mean to show that presence predictions from the deep learning architecture are accurate enough to estimate useful and more abstract perceptual indicators.

3. *Title - The perceived time of presence of sources is an intermediate goal, no? Much of this paper is focused on investigating pleasantness. Possibly change the title to match this goal*

   → The main objective of this work is the prediction of the perceived time of presence of sources of interest for the perceptual characterization of soundscapes. We study the estimation of pleasantness as an application, but the developed method could also be used to predict other quality descriptors or in mapping applications (eg. [1]). The structure of the paper has been rearranged to emphasize this aim.

4. *Section 2.1 - Each of the recordings from the 19 locations are 45 seconds long? Including the 6 recordings that weren't replicated? This could be made clearer.*

   → All 100 sound scenes (75 simulated, 19 replicated and 6 recorded) in the listening test corpus are 45 s long. For the replicated and recorded scenes, the 45 s extracts are parts of longer recordings. This is now more clearly stated in Section 2.2.

5. *Section 3.1 - The motivation and description of the computation of the time of presence approximation from the synthetic soundscapes could have been clearer. I think you should emphasize that this computation requires having separated tracks like in the synthetic soundscapes. And then also emphasizing the optimization of alpha and beta such that these measures match the perceptual measures as closely as possible. This all didn't quite make sense until the next section in which I realized they were going to be used as the training data for the ML model.*

   → These points are now more clearly explained in Section 2.4, which is dedicated to the evaluation of indicators to be used in the automatic annotation of the simulated dataset for deep learning.

6. *Section 3.1 - Was there any overall background noise, e.g. brown noise or "urban rumble", added to the recordings to lower the SNR to something more realistic?*

   → In this study no background noise was added in a controlled manner. However, some background noise is present in the isolated samples used to simulate sound scenes. Particularly, each scene contains at least one background source for which recordings obtained on Freesound are used. These recordings were selected to be representative of the source but were not cleaned through pre-processing, leading to (uncontrolled) background noise being present to some extent in each simulated scene.

7. *Section 3.2 - 125 ms is quite a large hop size. Since you are using synthetic data anyway (and thus are not very data limited), consider a smaller hop size, as is typical in most machine listening literature (e.g. 5 - 20 ms). Especially considering that you are detecting birds which are quite short sound events.*

   → Information on smaller time scales would indeed be more appropriate to detect sound events (from a few milliseconds for birds to tens of milliseconds for speech). Though, this work is placed in the context of a monitoring application through sensor networks in urban areas, for which using features computed on fine time scales may not be viable in the long term due to storage and privacy concerns. This motivates our choice to operate on the third-octave spectrum with 125ms hop size, which is typical in monitoring studies. This is now discussed in the introduction.

2 Answers to Reviewer II

1. *The abstract should be slightly revised; there is too much "background" info, while the reader might want to see already something about (quantifiable) results here.*

   → The abstract has been revised to reduce background information in favor of the description of the contributions and results of the current study.

2. *The first two paragraphs of the Intro might be re-thought, they are not very well connected and do not add much to the narrative to frame the issue under investigation. I would expect a more extensive literature review about sound sources recognition algorithms and auditory scenes characterization methods which is currently missing.*

   → The introduction has been rewritten to better introduce the objectives of this study, as well as related work in machine learning, monitoring applications, and soundscape perception.

3. *The claim that the 45-second excerpt is perceptually relevant should be supported by some evidence/reference somehow.*

→ Many studies dealing with the environmental acoustic quality are associated with duration of acoustic measurements ranging from few seconds for the shorter [2, 3] to fifteen minutes [4, 5] and even to eighty minutes [6]. Due to the limited duration of perceptual tests because of the fatigue, stimuli between 30 seconds and 1 minute are often preferred for laboratory tests.

4. *The process of manual annotation of the excerpts should be described in more detail to make it replicable.*

   → The 74 scenes in the reference corpus were obtained as part of the GRAFIC project [7] and manually annotated in [8]. These annotations are reused in this paper, as is now stated more clearly in Section 2.1.

5. *In general, references should be added when claims are made, it is not suitable to leave to the reader the task of tracking back such info. E.g., lines 176-180*

   → References have been added throughout the paper when applicable.

6. *Some more details could be provided about the procedure and technical equipment used for the perceptual experiment.*

   → The description of the equipment used in the listening test and its calibration has been extended in Section 2.2.

7. *When applicable, details of the ethical review process should be reported.*

   → Following ethical procedures, the participants were informed on this study via written information, and written consent was obtained for all of the participants. The present study does not require research ethics approval according to French law on ethical review of research involving humans. It does not concern specific health problem, any vulnerable groups or sensitive issues, it does not involve any method that aims to affect the participants physically or mentally, nor is it associated with any risks beyond those of everyday life.

8. *Some concepts are introduced but not always clearly defined. For instance - prominence (is it proximity/intensity?), emergence (is it signal-to-noise ratio?), etc. More aspects are mentioned in section 3, but the reader might want to know more already at this stage, and it would be desirable to discuss them more extensively.*

   → Concepts such as source dominance discussed in this paper are subjective descriptors of source activity, and as such have no clearly defined relation to the acoustical properties of the audio signal.

9. *In Section 5 I was longing for more discussion about possible challenges and issues one might face when the complexity of the auditory scenes increases (e.g., more sources), and when other environmental non-acoustic factors (e.g., match between auditory and visual settings) come into play, which could not be detected within the proposed framework.*

→ Section 5 now discusses the generalization and extension perspectives of the proposed approach in more detail. Its adaptation to new sound sources or environments is straightforward on the condition that enough data (both isolated samples of new sources and reference scenarios) can be gathered. As non-acoustic factors, visual settings are known to have an influence on soundscape perception and quality (see [9, 10, 11]). Though, these interactions go beyond the range of this study in the context of acoustic sensor networks where only sound is recorded.

3 Answers to Reviewer III

1. *I am not sure what to make of this study. I simply do not get the message, if there is one. The article seems to lack a clear scientific question, and consequently there is no answer - no message. [...]*
The arguments the authors present in the introduction for conducting this study do not convince me. Surely, future smart cities could have all sorts of sensors all over for monitoring various types of physical entities, including sound. But, would that motivate the present study and its approach? [...]
In general, I often find it hard to follow the presentation of the study and its results, and often get confused. Perhaps this is because there is no red thread in the story or no clear scientific question to answer. [...]
Overall, for this article to be successful, the authors must agree on one clear scientific question and one clear message. Then they must write a clear story, guiding the reader to the conclusions. Remember to answer your question.

   → The aim of this work is to determine wether a deep learning architecture can identify active sound sources in polyphonic urban sound environments from typical monitoring measurements (third-octave spectra). It is motivated by the fact that the composition of sound environments has a known influence on the perception of pedestrians, but is currently not explicitly studied in monitoring applications based on sensor networks. We achieve this goal by learning on labels approximating the perceived time of presence, which can be obtained in an automatic way if the ground truth source contributions in the training set is known. Pleasantness prediction is studied here as an application to assess the usefulness of the proposed approach.
   The structure of the manuscript has been revised to clearly expose the contributions of this work:

   - Section 2 details the deep learning dataset construction, its validation through a listening test and the choice of acoustic indicator for its automatic annotation,
   - Section 3 presents the training and results of the source detection deep learning model,
   - Section 4 applies this model to pleasantness estimation.

2. *At Inter-Noise in Hamburg in 2016 Lundén, Axelsson and Hurtig presented the results from a similar study, which seems to have been driven by the curiosity of learning whether machine learning is useful at all and what acoustical features would predict the outcome of soundscape assessments. This does not appear to be the motivation for the present study. Yet, the authors place a lot of energy on reporting on which regression model provides for the best predictions. Why is that?*
   Scientifically, the most interesting question related to this study is: what acoustical properties would be most relevant to monitor in the future if we are to replace the current sound level approach with one that is more holistic? It seems that the authors have some insight into this. I recommend refocusing the article on this question. It would make the article a lot more worthwhile. Leave the potential practical applications of the results for the discussion, rather than trying to use them as a motivation for the study in the introduction.

   → The original structure of the manuscript was misleadingly focused on pleasantness estimation, though it is in fact only an application of the proposed approach studied here to evaluate the quality of source presence predictions. As such, the motivation of this study is entirely on the identification of active components of the soundscape, which could have several fields of application from quality assessment to information-rich mapping.

3. *On Rows 39-40 in the introduction the authors claim that noise maps stipulated by the European Environmental Nose Directive (2002/49/EC; please correct the typo on Row 37) would provide "useful tools for urban planning and noise reduction purposes." I must say that urban planning is not at all on the EC agenda in this respect. The purpose of the noise maps is to provide intelligence for noise actions plans only. Urban planning would require a completely different approach than what the present noise abatement framework provides.*

   → The claim that the European Commission 2002/49/EC directive was aimed at urban planning has been removed from the introduction.

4. *The quote related to ISO 12913-1 on Rows 43-45 is incorrect. Please correct.*

   → Following the changes in the introduction structure, the quote related to ISO 12913-1 no longer appears.

5. *In particular, it is hard to understand the presentation of figures, tables and equations. They must be more fully explained in the text.*

   → Figures, tables and equations are now more clearly explained in the text.

6. *I recommend the authors to occasionally remind the reader about the meaning of the different abbreviations used. Now I must move back and forth in the text to try to remember what they mean, which causes me to lose track*

of where I am going. This makes it very hard to interpret and to understand the results. It is necessary to support the reader in understanding the text.

→ Several abbreviations have been removed or explained throughout the paper to improve its overall readability.

7. *Also keep in mind that the reader may not be an engineer, physicist or mathematician and may not understand the equations. Please explain the equations in plain language, and help the reader to understand what they mean. For example, what does the second derivative of a function mean (velocity or acceleration)?*

   → Explanations for the main equations in the manuscript have been added (eq. 1-3). Specifically, in eq. 1 the derivation is applied once on each dimension of the third-octave spectrum (hence the term second derivative), which underlines variations in time and frequency.

8. *I fail to understand how the different parts of the study relate to each other. How does the perception study relate to the machine learning part? Are they connected, or are they completely separate? I also got confused when I learned that there were some 400, or was it 200, simulated acoustic scenes used for machine learning, while there first were some 75 such scenes.*

   → The structure of the paper has been changed to better link different parts of the study. Section 2 presents the generation of a large simulated dataset with 400 sound scenes for training a source detection model and an additional (independent) 200 scenes for its evaluation. A third corpus containing 100 sound scenes including 75 simulated, 19 replicated and 6 recorded, is constructed in Section 2.2 as part of a listening test to assess the change in perception in simulated scenes and to obtain subjective evaluations useful in the rest of the paper. In Section 2.3, the listening test results motivate the choice of the indicator used to automatically annotate the perceived time of presence on the 600 total scenes used in the machine learning approach.

9. *The Equations 5-11 must be explained in more detail. Please also inform the reader that you created multivariate linear regression models and how this was done. It was not until I saw the equations that I understood that they are regression models. This is not explained in the text. In general, you must present to the reader how statistical analyses are conducted and with what statistical package. At present, there is no such information in the text.*

   → Additional explications were added to equations 5-11 (now 5-8), and the procedure used to obtain baseline models from perceptual and acoustical parameters through multivariate linear regressions has been detailed in Section 4.1. All statistical analyses are conducted using Matlab and its statistical toolbox (code is made available on the companion website) and

the machine learning part is implemented in the Pytorch framework (also available). This is now stated where applicable.

References

[1] C. Lavandier, P. Aumond, S. Gomez, and C. Dominguès. Urban soundscape maps modelles with geo-referenced data. *Noise Mapping*, 3:278–94, 2016.

[2] R. Paulsen. On the influence of the stimulus duration on psychophysical judgement of environmental noises taken in the laboratory. In *INTER-NOISE 1997*, 1997.

[3] G. Brambilla and L. Maffei. Responses to noise in urban parks and in rural quiet areas. *Acta Acust. unit. Acust.*, 92:881–6, 2006.

[4] S. Kuwano, J. Kaku, T. Kato, and S. Namba. The experiment on loudness in field and laboratory: An examination of the applicability of laeq to mixed sound sources. In *INTER-NOISE 1997*, 1997.

[5] B. De Coensel, D. Botteldooren, B. Berglund, M. Nilsson, T. De Muer, and P. Lercher. Experimental investigation of noise annoyance caused by high-speed trains. *Acta Acust. unit. Acust.*, 93:589–601, 2007.

[6] S. Namba and S. Kuwano. Measurement of habituation to noise using the method of continuous judgment by category. *J. Sound Vib.*, 127:507–11, 1988.

[7] P. Aumond, A. Can, B. De Coensel, D. Botteldooren, C. Ribeiro, and C. Lavandier. Modeling soundscape pleasantness using perceptive assessments and acoustic measurements along paths in urban context. *Acta Acust. unit. Acust.*, 103:430–443, 2017.

[8] J.R. Gloaguen, A. Can, M. Lagrange, and J.F. Petiot. Creation of a corpus of realistic urban sound scenes with controlled acoustic properties. In *Meetings on Acoustics*, 2017.

[9] S. Viollon, C. Lavandier, and C. Drake. Influence of visual setting on sound ratings in urban sound environment. *Applied Acoustics*, 63:493–511, 2002.

[10] P. Ricciardi, P. Delaitre, C. Lavandier, F. Torchia, and P. Aumond. Sound quality indicators for urban places in paris cross-validated by Milan data. *J. Ac. Soc. Am.*, 138:2337–2348, 2014.

[11] R. Pheasant, K. Horoshenkov, G. Watts, and B. Garett. The acoustic and visual factors influencing the construction of tranquil space in urban and rural environments tranquil spaces-quiet places? *J. Ac. Soc. Am.*, 123:1446–57, 2008.