Reply to reviewers concerning submission AAA-D-19-00080R1: "Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques"

November 29, 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, which are summarized here. The next section list our answers to each of the reviewers comments, with references to the revised manuscript where appropriate.

1 Answers to Reviewer III

1. *The revised version of this manuscript reads somewhat better than the original version. However, it still would benefit from further restructuring and the provision of further details. In the response letter, I find a lot of detailed information in the authors responses to the reviewers that are not implemented in the manuscript. If one think that some information is valuable to the reviewers, it is equality relevant to the readers of the published work. In general, all information provided to reviewers in the response letter must be implemented in the manuscript. For many journals, failure in doing so is reason enough for rejecting a manuscript from being published. Please, revise the manuscript accordingly.*

   → All elements provided to the reviewers as part of the previous response letter are now presented in the manuscript. In particular, the choice of 45 s long extracts for the listening test and absence of controlled background noise in simulated sound scenes are now included in Section 2.1, on rows 255-262 and 204-207 respectively.

2. *As I understand this manuscript, it contains three parts. Consequently, I propose that each part of the study should be allocated its own separate section.*

Section 2 should report on the listening experiment, only. I propose that it is structured into subsections: 2.1 Participants, 2.2 Stimuli, 2.3 Data collection instrument, 2.4 Equipment, 2.5 Procedure and experimental design, and 2.6 Results. Add or remove subsections as appropriate.

→ Section 2 has been rewritten according to the following structure: 2.1 Stimuli, 2.2 Equipment, 2.3 Participants, 2.4 Procedure, 2.5 Results.

3. *Section 3 should report on the machine learning study, including the corpus used and how it was generated. I still do not understand how the latter was done. How many different sounds for each source (traffic, voices, birds) were used to generate the scenes?*

   → The machine learning corpus is now described in Section 3.1 (instead of Section 2.1). As is now emphasized, its generation procedure is the same as for the simulated sub-corpus of the listening test. However, the isolated samples database is different. Details on the contents the isolated samples database have been added in Section 2.1 (rows 197-204) and 3.1 (rows 508-512) for the listening test corpus and the deep learning corpus respectively.

4. *For the reader to be able to evaluate the validity of the study, it is necessary to know some more about how simScene works and how it utilized the information from the 74 recordings from the GRAFIC project to generate the scenes used.*

   → The description of the generation procedure from distributions extracted from the GRAFIC project recordings using simScene has been extended in Section 2.1, with a presentation of simScene simulation modes and inputs on rows 162-193.

5. *Section 4 should report on the application of the machine learning model to the prediction of pleasantness. The purpose and objective of this section must be clearly and explicitly communicated to the reader at the beginning of the section, providing a clear focus. Currently, the purpose of the section is not clear enough.*

   → The objective of Section 4 is to present a potential application of the proposed detection model to the prediction of higher-level perceptual attributes, and assess the gain of such an approach with respect to known models. The introduction of Section 4 has been extended to better present this motivation.

6. *Probably the presentation of the acoustic indices (now in Section 2.4) should be moved to Section 4, where they are used.*

   → Acoustic indices are first used to identify a potential automatic annotation of the deep learning corpus in terms of perceived source presence. The presentation of these indices has thus been moved to Section 3.2.

7. *In describing what the authors did, they must be precise and consistent in the terminology. For each phenomenon described, they must use exactly the same terms throughout the manuscript without any variation, so that the reader may understand what they are referring to. At present, there is too much variation in expressions, which is confusing. For example, on Row 261 the authors mention 75 selected extracts. What are those? Are they the same as the 75 simulated scenes mentioned previously?*

   → The terminology used throughout the manuscript has been revised and is now more consistent.

8. *For the listening experiment, I understand it as if the authors used 100 sound scenes, out of which 75 were simulated, 6 were recordings and 19 were replicated scenes from a previous project. What does it mean that the 19 scenes from a previous project were replicated? Where they simulations that match previously recorded scenes, or were they new recordings that are similar to previous recordings?*

   → As is now more clearly described in Section 2.1 (rows 292-299), replicated sound scenes are generated using simScene, with scene composition (scenarios) constrained by annotations of the reference recordings. This differs from the 75 simulated sound scenes for which the scene composition is obtained by sampling from gaussian distributions.

9. *On Row 247, the authors mention that 200 scenes were generated. Does it mean that the 75 simulated scenes used were selected among the 200?*

   → As the generation procedure for simulated scenes includes random sampling from distributions, simScene may output similar sound scenes. To avoid this phenomenon in the listening test corpus, 200 scenes are first generated, among which 75 scenes are selected to maximize content diversity. This is described on rows 262-274 of the revised manuscript.

10. *On Rows 256-260: The authors write that "Ideally, sound scenes should cover the resulting 3-dimensional space in a homogeneous way. To do so, the 75 scenes that maximize the minimum pairwise distance in the 3-dimensional space are selected."*
    I am not sure that I understand what 3D space the authors are referring to. Do they mean the 3D space of Traffic, Voices and Birds presented in Figure 1? If so, then please spell this out explicitly in the text, including what this 3D space represents and how it was created. Please also explain how to "maximize the minimum pairwise distance."
    I find it very hard to understand Figure 1 from the limited information provided. It must be improved (or removed).

    → The dimensions of the mentioned 3D space correspond to the presence (0-100%) of traffic, voices and birds sources. The 75 simulated scenes of the listening test corpus are selected to be the most isolated, that is distant from others with regard to the euclidean distance, from the original pool of 200 simulated scenes. The selection procedure description has been rewritten on rows 262-274 and Figure 1 has been removed as it did not directly contribute to this description.

11. *On Row 285, I wonder what it means that the authors "normalized" the sounds. Does it mean that the authors calibrated the sounds to the authentic sound levels?*

    → The sound scenes are normalized so that they are heard at the desired sound level (as measured in the case of recorded and replicated scenes, or as sampled conditionnally to the ambiance in the case of simulated scenes). This requires the relation between the electrical level at the output of the sound card and the resulting sound level produced by the headphones, which is obtained in Section 2.2.

12. *On Rows 290-291, the authors write that the participants used 8 criteria to evaluate the sounds. I suppose that it means that they used 8 attributes. It also seems as if the authors used 11-point bipolar scales, delimited by opposite terms. Then the term 'Likert scale' used on Row 370 is incorrect. A Likert scale is a 5-point category scale asking to what extent a person agrees or disagrees with a statement. As far as I can tell, the authors did not use any Likert scales.*

    → The term "Likert scale" has been replaced with "semantic differential rating scale" throughout the manuscript.

13. *Rows 329, 345, 352 and 414: Replace 'subjects,' with 'participants.'*

    → All mentions of "subjects" are now replaced with "participants".

14. *I do not understand the information provided in the paragraph about the sound level calibration on Rows 334-348. Particularly, I do not understand the rationale for the calibration procedure. It must be explained in some more detail. Perhaps a conceptual diagram showing how the different equipment were connected would be helpful? The reader should be able to replicate what was done based on the information provided in the manuscript.*

    → From the calibration procedure, a simple (linear) relation is obtained between the voltage at the output of the sound card and the sound level produced by the headphones. Sound scenes can be normalized using this information so that they are played at the desired sound level using the specific headphones and sound card configuration. The steps undertaken during the calibration procedure are now detailed in Section 2.2 (rows 315-347).

15. *On Rows 401-403 it seems as if the 6 recorded and 19 replicated scenes (n=25) all were recordings. Is that so?*

    → The 19 replicated scenes are generated by simScene, and thus are not recordings. This part has been rewritten to avoid any confusion in this regard (rows 543-455).

16. *Figures 2 and 3: Please observe that these two figures present Principal Components Loadings for the five attribute scales. Because the scales are bipolar, it is confusing to see lines extending in two directions. There should only be five vectors extending from the origin of the space to the points in the 2D space that represent the coordinates of the loadings on the two Principal Components. Moreover, I am not sure that it is necessary to include the ellipses, short arrows and datapoints in these figures. I just find it messy.*

    → Both figures displaying the principal components analyses have been modified so that loadings are in one direction only. Furthermore, the ellipses and arrows have been removed along with the corresponding text. They are available as supplementary material on the companion website.

17. *How were the PCA conducted and with what statistical package? [...] What statistical package was used to calculate the correlation coefficients? [...] What statistical package was used for conduction the regression analyses?*

    → All statistical analyses are conducted using Matlab R2015b with the Statistics and Machine Learning Toolbox v10.1. This is now emphasized in Section 2.5, and reminded where appropriate. The code is also made available for further information about the specific functions used.

18. *Equations 2 and 3 must be better explained, and presented separately. Please use plain language, to explain what the equations mean.*

    → Both equations are now presented separately, and each one is more thoroughly explained.

19. *Table 2: I expect to see * p¡0.05 and ** p¡0.001 as foot notes below the table and not in the caption.*

    → Significance thresholds have been moved to table footnotes where applicable.

20. *Are they 1- or 2-tailed probabilities?*

    → The Matlab *corrcoef* function was used to compute all correlation coefficients. The corresponding significance test is two-tailed. This is now stated in Section 3.2 (rows 632-635) and 4.1 (rows 666-669), where correlation coefficients are presented.

21. *The authors must also explain all abbreviations in the table. It must be possible to read tables and figures independently of the text body. Please, include all correlation coefficients in the table, also those that are not statistically significant.*

    → Abbreviations of perceptual attributes have been added to the respective table captions. Non-significant correlations have also been added to Table 2.

22. *In Section 4, I miss the relevant F- and t- statistics for the regression models. For each model, one must report the overall F-statistics and the t-statistics for each factor included in the models. This is typically presented in the text body.*

    → F- and t- statistics have been added to the text following equations 5 and 6 on the baseline models of pleasantness.

23. *For each model, the authors must also explain what data was used (which corpus).*

    → Baseline models of pleasantness are computed on the listening test corpus, as the considered perceptual attributes are available on this corpus only. The baseline model from acoustic indices is computed on a subset of 92 sound scenes as the computation of time of presence estimations requires separated source contributions, which are not available in the 6 recorded scenes. These elements are now more clearly described in Section 4.1 (rows 783-787 and 809-817).