Reply to reviewers concerning submission sensors-238686: "An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach"

November 20, 2017

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 I

1. *It is recommended to discuss about noise annoying and loudness monitoring and evaluation with sensor network.*

   → This is now more clearly discussed in the introduction as well as in the conclusion where a dedicated paragraph has been added on the use of this scheme as part of the current project.

2. *Is it possible to achieve a two channel coding and transmission with this proposed scheme?*

   → While a single audio channel was selected for this study, a multiple channel encoding can be performed by applying the Delta compression along both the time and channels dimensions. As the transmitted data primarily consists of energetic measurements, additional channels are likely to be encoded very efficiently due to the redundancy of information already contained in the first channel. Future work using this idea in a sensor grid is now discussed in the conclusion.

3. *Typos:*
   - Small English typos.

- Use of "we" too often. It is recommended using third person writting.
- Equation numeration

→ The typos have been corrected and the whole manuscript has been carefully scrutinized. The use of "we" has also been reduced to favor third-person writing. The equation numeration has been corrected throughout the document.

2 Answers to Reviewer II

1. *The Introduction is a bit shaky and needs to be reworked in my opinion. This section should: a) provide a literature review of the state of the art on acoustic sensors networks; b) identify a methodological gap in current practice; c) explain why it is important to fill it; and d) give a glimpse about how the proposed research aimed to do so. The current structure is confusing to me, with Section 2 coming a bit out of the blue (I feel it should be reported earlier in the Introduction).*

   → Following your comment as well as the comment of Reviewer III, we have rewritten the introduction to better highlight the contributions of the paper. The considerations on acoustic monitoring and event detection in urban sound environments previously discussed in Section 2 have also been moved earlier in the introduction to better frame those contributions. A new section (Section 2.1) was added to discuss the choice of data representation.

2. *Please, also note that not all readers are likely to be familiar with ASC and AED protocols, so it would be nice if the authors could provide some more details about the general steps; i.e., feature extraction (parametrisation) stage, classification stage, etc.*

   → We added a new figure (Figure 6) to clarify the applied classification process.

3. *The Validation protocol (Section 4) is not very convincing. In particular, the subjective validation of Section 4.3 does not look rigorous at all.*

   → Our coding scheme must observe several constraints, namely low bitrate, measurement accuracy for monitoring purposes, the ability to perform event classification on the transmitted data and inintelligibility for privacy concerns. For each of these required properties we designed a validation procedure using urban environmental sound recordings. We acknowledge that the listening test is somewhat informal, as our goal is more to provide insights about why the transmitted data will not lead to intelligible speech content than to fully prove it which would be out of the scope of this paper.

4. *In general, a strange aspect for me is that, this paper being submitted to the journal Sensors, there is no actual description of the sensors network*

architecture and/or hardware components proposed for the current study.

→ The main contribution of this work, which is now more clearly shown in the introduction, is about the definition of a coding scheme that will be used (as stated in the conclusion) in a sensor network that will be implemented in a near future. But prior to this usage, several properties needed to be validated, motivating this work.

5. *The overall structure of the paper should be re-thought in my opinion, as the reader struggles to orientate himself between the different sections of the manuscript. When I read the conclusions, it is hard to track back the discourse in the methodological and results sections.*

   → We added more references between sections to facilitate the reader's orientation in the manuscript.

3 Answers to Reviewer III

1. *It would be nice if the authors could add the following information to their work: Motivation for using the third octave band spectral representation.*

   → A new section (Section 2.1) has been added to motivate the use of third-octave bands for both acoustic monitoring and audio event detection.

2. *Clearly state the novel points of this work in the introduction.*

   → The contribution is now more clearly stated in the introduction.

3. *Why did you select these classification schemes? There many approaches in the literature exploiting that; for example "A novel holistic modeling approach for generalized sound recognition". In addition to that, the authors should justify why the did not employ deep learning (for both feature extraction and pattern recognition).*

   → As is now explained in Section 3.2, we do not aim at achieving the best classification performance but 1) to prove that third-octave bands match the most commonly used descriptors for the classification task, and 2) to discuss the effect of encoding parameters on the recognition accuracy to select the most appropriate. To this aim we chose to replicate the baseline results for the considered dataset on these four schemes. The study of more advanced classification schemes and of deep learning for AED is very interesting and will be of particular interest in future work.

4. *Another point regarding detection would be to relate this work with other operating in real unrestricted environments, such as "Acoustic detection of human activities in natural environments" and how the proposed method could boost the existing ones.*

   → This would be a very interesting avenue of research that could be considered for future work. The discussion on data representations and

models used in environmental audio event recognition (Section 2.1) has been extended to include related works.

5. *Since the authors use the UrbanSound8k dataset, it is not clear how the sensor grid approach is applied. Is there a grid of sensors included?*

   → As now more clearly stated in the introduction and conclusion, the proposed scheme will be applied to a sensor grid as part of the CENSE project. However, as this sensor grid is not yet implemented, the UrbanSound8k dataset provides both a decent amount of audio data with similar content (urban environmental sounds) and classification methods and results that can be used as a baseline for our work.

6. *The authors should compare their system with other existing in the related literature to understand its merits and limitations. This should be done in many levels, i.e. coding, recognition, detection, etc.*

   → The contribution of this work is on the coding only. As discussed in the introduction, to the best of our knowledge the proposed approach is the only scheme where the data transmitted is not the detected events nor the raw audio but an intermediate spectral based format, specifically designed for acoustic monitoring and AED, thus motivating this paper.

7. *Regarding intelligibility, it would be nice if the authors gave more information on the used metric. Moreover they should use standard metrics used for that which are similar to those used in the speech enhancement field. See paper "Objective comparison of speech enhancement algorithms under real world conditions" for more details.*

   → While we are aware that standardized metrics exist such as the ITU protocols, we found that they are not adapted to the evaluation of intelligibility in this application. The encoding process distorts the signal in a very non-linear way, so that we have no control over the background and signal elements. A five-point score on the overall intelligibility is still used however, and a second metric (the transcription ratio) is designed to account for the typical distortions induced by linearly scaled spectrogram and phase recoveries. Motivation and insights concerning the used metrics as well as this discussion are now added in Section 3.3.