Reply to reviewers concerning submission JASM-D-18-00019 "Relevance-based Quantization of Scattering Features for Unsupervised Mining of Environmental Audio"

August 17, 2018

As a preamble, we would like to thank the reviewers 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 Answer to Reviewer 1

1. *I suggest that they only refer to ecoacoustics in the discussion by mentionning that this could be tested on other audio scenes as those recorded in natural environments.*

   → The abstract introduction, background section and conclusion have been rewritten to widen the scope of application of the proposed method and suggest the ecoacoustic as potential area of interest for further research.

2. *The dataset is too quickly described. I had to read ref#36 to understand how it was built and made of. It would be nice to have an example of a labelled recording either as a spectrogram or scalogram to estimate the complexity of the scenes recorded. The event classes should also be detailed.*

   → We have listed the classes from the DCASE taxonomy in full in Section 6.1. In addition, we have included the scalogram of a labeled recording (*park04*) as Figure 3.

3. *It is a pity that the methods are not directly compared with the different strategies reported in ref#36.*

   → Since the tasks are different, it is not straightforward to compare the proposed methods with the approaches of ref#36. In the former, we pro-

pose a method for acoustic scene similarity retrieval, while the latter discusses the problem of scene classification. In similarity retrieval, we would like to find other recordings in the dataset that are as similar to the original as possible (which here is defined as being in the same class). In scene classification, the task is to assign a class to each recording. In this sense, similarity retrieval is a more general problem. For our experiments, we calculate the p@k curves for each proposed method, allowing us to analyze the distribution of recordings most similar to a particular query recording. The classification setting is reproduced for $k = 1$, with the p@1 measure corresponding to the leave-one-out classification accuracy of nearest neighbors. As such, the classification task is a special case of the similarity retrieval setting considered here. We have modified the manuscript to make this distinction more clear.

It is also worth considering that the BOF model is used differently in scene retrieval (our use case) and scene classification (ref#36). In scene retrieval, given two scenes (a and b) modeled using GMMs (A and B), the similarity between two scenes is the similarity between A and B. In scene classification, each class of sound scenes (park, boulevard) are modeled using a GMM (P, B). Then, for an unknown scene u, the likelihood of u given P and B are computed. The scene u is then labeled P if the likelihood of u given P is greater than the likelihood of u given B.

Even if the technical aspects are quite close, the way the algorithm is designed is different as the tasks they have to solve are differents. A synthetic description of the matter is added to Section 2.

4. *.P1L36: "superiority" and other superlative seems too strong. I would say that the proposed approach performed better on the dataset tested.*

   → VL ✓

5. *.P2L29: "cannot be trusted" seems too strong, any system is prone to errors. These errors need to be estimated so that we can estimate how much the system is reliable, but I would not roughly say that it cannot be trusted.*

   → VL ✓

6. *.P10L44: the title of this subsection does not seem to be totally appropriate, I would rename it "Evaluation and algorithm"*

   → VL ✓

7. *.P11L37: the comments about figure 3 (and others) is rather quick with no try of quantification of the differences/similarities between the p@k curves.*

   → This section has been expanded in the revised manuscript to provide more interpretation of the results.

8. *.P12L37: Delete the comment addressed to one of the authors*

   → This has been fixed in the revised manuscript.

9. *.P13L16: Using French for a project name might not be a good idea for reaching a large audience*

   → This has been fixed in the revised manuscript.

2 Answer to Reviewer 2

1. *Only two applications are considered in the literature review, bioacoustics and urban sound environment, but they allow to present a problem that also exists in other application areas.*

   → We reformulated in a more abstract way the task we attempt to solve, and give potential areas of applications in the introduction and conlusion sections.

2. *The authors present their technique as opposed to the BOF approach, that is considered "state of the art". BOF maybe a popular technique but the fact that gets rid of the temporal structure makes it useless in many tasks. In fact, in DCASE 2013 BOF was the chosen baseline system but no participant proposed that technique. Certainly, BOF allows the authors to make stronger the point about describing an acoustic scene from a few distinct events, but perhaps they should avoid to claim BOF as "the" state of the art in this problem.*

   → As discussed above in the response to the first reviewer, the application in the DCASE 2013 paper, classification, is different from the one proposed in our work, similarity retrieval. For classification, the BOF approach is indeed not the state of the art. However, in acoustic scene similarity retrieval, we are not aware of any systems which outperform the BOF. As such, it is indeed the state of the art for the similarity retrieval task. The manuscript has been revised to clarify this point.

3. *It would be interesting to see the contribution to the results of the Gammatone wavelet in comparison to Morlet or another more classical wavelet.*

   → VL ✓ We have not re-run experiments with Morlet wavelets instead of Gammatone wavelets, for two reasons. First, Morlet wavelets do not fit the whole narrative of the paper as well as Gammatone wavelets: while the arguments in favor of Morlet wavelets arise from harmonic analysis (Heisenberg time-frequency uncertainty tradeoff), those in favor of Gammatone wavelets arise from auditory neurophysiology. Because we work in the context of computational auditory scene analysis, *a fortiori* in an unsupervised setting, we believe that it is preferable to borrow from a well-studied auditory model (Gammatone) rather than from a general-purpose operator (Morlet). The second reason why we did not conduct a systematic benchmark between Morlet and Gammatone is that, in previous experiments involving supervised classification of musical instruments with scattering coefficients, Gammatone wavelets appeared to outperform

Morlet wavelets by a small margin, from 80% and 79% test accuracy respectively. In all likelihood, the effect of the specific choice of wavelet is also small in the case of unsupervised similarity retrieval between auditory scenes, and possibly not statistically significant given the limited sample size of the DCASE 2013 dataset. We have expanded Section 3.4 to highlight these considerations. In addition, we have included a mathematical description of the Gammatone wavelet filter bank employed in this study as supplementary material. Indeed, the design of our Gammatone wavelet is slightly different than in the reference papers of Venkitaraman and Adiga, because, unlike them, we need to account for the influence of both the quality factor Q and the polynomial degree parameter N. It is our hope that the proof of the formula linking Q and N to the parameter α in the Gammatone function will be beneficial to the signal processing community at large.

4. *There are several parameters that must be set in the experimentation. The reader may want to know if there is any specific parameter that is critical.*

   → Considering the features design, we set the parameters to widely used values that have shown to be robust. We also ran a sensitivity analysis for the parameters that controls the proposed algorithm: scaling parameter q and the number of clusters M. None of them are found to be critical providing that they are set in a reasonable range. It is worth noting that the scaling method is specifically designed for reducing sensitivity compared to the direct RBF kernel formulation.

5. *What about M?*

   → The number of clusters M is indeed an important parameter. The higher M is, the more complex the model and the computation are. Setting M to 4 roughly states that there may be at most 4 different types of events in the given chunk of audio, here 30 sec. The analysis we added to the manuscript at the end of Section 7.

6. *According to Section 6.2, the number of scattering features is much higher than MFCC features; 1367 instead of 40 for each vector, and there is a similar number of vectors per time unit. This would mean a much higher computational load. However, in the experiments, a projection is applied to reduce them to 30. Is it PCA? The convergence of the EM algorithm is invoked, but the computational load may be an added reason.*

   → As described in Section 6.3, the scattering features are indeed projected onto the top 30 principal components for use in the BOF model. This is to mitigate the problem of running EM in such a high-dimensional space. For the other methods, it may reduce the computational load, but potentially at the cost of sacrificing certain information that may be leveraged by those other methods. For this reason, we have chosen to run them on the full, 1367-dimensional, feature vectors.

7. *I do not see the authors have mentioned in Section 7 the following observation (Fig.3): for MFCC the relevance-based quantization improves the performance wrt BOF.*

   → The manuscript has been revised to include this observation.

8. *Pag. 2, line 48: what "this" refers to?*

   → This has been clarified in the revised manuscript.

9. *Pag. 2, line 49: "More closely matches... than...*

   → This has been clarified in the revised manuscript.

10. *Pag. 3, line 22: an cluster -> a cluster*

    → This has been clarified in the revised manuscript.

11. *Pag. 3, line 23: Unconsistent sentence "These clusters to define..."*

    → This has been clarified in the revised manuscript.

12. *Equation (5): Notation is different from previous equations*

    → The notation has been fixed for consistency with the other equations.

13. *Pag. 7, line 9: S1 should be S2*

    → This has been fixed in the revised manuscript.

14. *Pag. 9, line 48: suppress "the"*

    → This has been fixed in the revised manuscript.