Response to Reviewers: T-ASL-05872-2016 "Object-based Auditory Scenes Similarity Retrieval and Classification With Wavelet Scattering"

December 21, 2016

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 sections list our answers to each of the reviewer's comments, with references to the revised manuscript where appropriate.

1 Responses to the Associate Editor

1. *Line 40, α may appear in previous equations*

   As far as we can tell, α only appears during the definition of the $\widehat{\text{EMD}}$ distance and not anywhere else.

2. *The argumentation toward the log compression is well conducted, and meaningful in this issue. The authors may simply swap the order of Fig. 3 and 4.*

   The order of the figures and their corresponding text have been reversed.

3. *The class accuracy Table III for Home and Residential area are singular. Authors shall discuss that.*

   The text has been modified to discuss these results.

4. *Complete caption of Fig. 5.*

   The caption have been completed to help the readability of the figure.

2 Responses to Reviewer 1

1. *Object-based is inappropriate, This paper introduces: no this paper uses the scattering network to try to tackle a specific sound classification task.*

The title has been changed to remove the "object-based" formulation. The particular approach introduced in the paper is the hierarchical representation of a scene as a partition of its distribution of scattering features. While the scattering transform, which has been introduced by previous works, forms the low-level layer of this representation, the hierarchical structure is novel and we believe generic enough to represent auditory scenes for a variety of information retrieval tasks.

2. *The second paragraph of the abstract is vague.*

   This paragraph has been reworded and worked into the rest of the abstract in a more fluid manner.

3. *No novel technique is presented.*

   Again, the hierarchical representation proposed is, to the best of our knowledge, novel and we provide adequate evidence, in our opinion, for its power in the proposed similarity retrieval task in the case of environmental audio data processing.

4. *The whole paper architecture is poor.*

   We have reworked the paper as a whole, trying to clarify the overall message. In particular, we reorganized the paper towards a unique claim, that is supported by experimental evidence. the contribution now explicitly target the acoustic similarity retrieval task. Complementary experiments in the more widely used classification task are used to validate the first step of representation using the scattering transform.

5. *All the justifications arise from psychoacoustics, bioacoustics ... when the challenge is to classify ambient non natural stationary sounds. For example, the whole Gammatone wavelet justification over Morlet is biased.*

   While some of the sounds are "non-natural", that is human-caused, that does not make differentiating between them an impossible task for a biological system evolved under exposure to other, more "natural" sounds. In addition, the psychoacoustic results cited have been obtained by conducting experiments where subjects listened to the same type of "non-natural" sounds. As such, we feel that the justifications are well-adapted to our method and tasks.

6. *The classification accuracy is not suited for comparison since the dimension of the feature vectors are not the same for all the techniques.*

   While the dimension of the features vectors are different, this does not necessarily have to impact the performance of the classification. For example, if we were classifiying stationary Gaussian signals, a spectral envelope feature, like the MFCCs, would prove to be just as accurate, if not better, than the richer second-order scattering transform. The important relationship to consider is whether the feature transform provides an accurate description of the relevant information content in the signal. In addition,

most of our experiments deal with the feature vectors through an RBF kernel, which is equivalent to considering the features embedded in an infinite-dimensional space, so in this sense the original dimensionality of the feature vectors carries less importance. A paragraph has been added to discuss the influence of dimensionality.

7. *The used classifier is not the same (from unsupervised GMM to supervised SVM)*

   We have included results for MFCCs using a linear SVM. However, we point out that the GMM is also supervised, although it is trained using a generative objective as opposed to a discriminative one, which is the case for the SVM.

8. *The paper reads like a recipe for the specific task without justifications, insights or proofs of the different given definitions and statements.*

   We justify the cluster-based representation with results from cognitive psychology. The use of the scattering transform is justified with invariance conditions of the tasks. Where necessary, the proofs and evidence are provided in references. The lack of specificity of this point makes it hard to address, but as stated above, we have reworked the whole paper to hopefully make our argument clearer.

9. *The dataset is not well presented since the classes seem to represent stationary noises such as car/tram/train*

   While these scenes definitely have a stationary component, such as for the car, tram, train, metro station, and bus classes. This is not sufficient to classify them, since this stationary signal is very similar for these classes. This is why the cluster-based hierarchical representation that we propose is necessary in that it allows the system to pick out the distinct non-stationary components that discriminate between the classes.

10. *"Recordings may be shifted locally without affecting their perception and therefore such shifts do not convey any information about the class" this is false. Rather say that application of time shift should not change the representation since it doesn't change the class belonging in this context and thus imposing time shift invariance will improve robustness.*

    That was indeed our meaning, the text has been changed to reflect this.

11. *The used descriptors are only locally time shift invariant and not globally which is never pointed out.*

    As formulated, the scattering transforms do indeed have only local time-shifting invariance, which has been clarified. However, the processing corresponding to "early" integrating achieves more invariance through complete averaging of the scattering transform along time and so does have that global time-shifting invariance.

12. *The function ϕ is hardly considered as a low-pass since a low-pass is only used in coordination with a high-pass for quadratic mirror filters. Here ϕ is a scaling function.*

    The function ϕ is very much a low-pass filter with a frequency support of approximately $1/T$ and a time bandwidth of T, where T is of the order of hundreds of milliseconds. It does not correspond to the traditional scaling function used in quadrature mirror filters.

13. *Line 34: "wavelet transform modulus" is not clear at all.*

    We have updated the text to clarify this expression.

14. *It is then explained that the features are stable to time warping yet the justification is for pitch variations which are frequency-warping*

    The justification of pitch variation is indeed a frequency-warping, but can be written as a time-warping. Indeed, a signal that is slightly compressed or dilated locally will have its pitch transposed up or down.

15. *The scattering network is defined with an infinite cascade of transform yet it is enough to compute two layers, why?*

    For the time scales considered in this paper, the first two orders are sufficient to describe the sound events present in the signals. This has been made clearer in the text.

16. *It is never specified what are the signals frequency sampling which we have to deduce to be 44100 Hz from line 9 page 3 "20kHz close to the Nyquist frequency of the audio recordings".*

    It is true that the audio recordings from the DCASE 2013 and 2016 datasets are sampled at 44100 Hz. This information has been added in Section VII.A.

17. *"Wavelet scalogram" is used along all the paper but scalogram is only defined for wavelets since it is a scale-o-gram.*

    This is true, but for readers less familiar with the terminology, we have opted to use this more redundant expression to express the fact that these derive from wavelet decompositions.

18. *Line 60: auditory and modulation filter nonsense.*

    We have modified the sentence to clarify our meaning.

19. *"unsupervised classification" it is clustering then.*

    The sentence have been rewritten to be more explicit.

20. *"time-frequency perspective or a machine learning perspective" is it imposed versus learned representation?*

    The intention was to situate the feature processing step either as post-processing of the scattering features or as pre-processing for the classifier stage. We have removed it during the revision.

21. *Equations 9 and 10 are superfluous.*

    These equations were removed from the paper.

3 Responses to Reviewer 2

1. *The late versus early integration of the features is well presented and evaluated. But a graphic representation of the whole process would arise the novelty of the model.*

   We acknowledge that such a diagram would help the reader, we are facing page limit constraints due to the extensive experiments described in the paper.

2. *Authors may precise how the Gammatone wavelets allow to reproduce the asymmetry of temporal masking and the asymmetry of spectral masking.*

   We have written a short explanation of the role of Gammatone wavelets in the modeling of the asymmetry of temporal masking and the asymmetry of spectral masking.

3. *Setting of Q_1 shall be motivated.*

   This has been added to the text.

4. *Reproducibility: data sets are publicly available, the author do not mention if they plan to make their model available.*

   Experiments have been performed in the MATLAB environment using open source libraries an all the authors are committed to comply to the reproducible research statements. The replication code is now available in beta, and will be soon stable.

5. *In the last experience (Table III) GMM (MFCC) may be compared to SVM (MFCC) (in addition to SVM (logscat)). There is somewhere a high dimension effect that shall be clarified.*

   We have added results for SVM on MFCCs. As mentioned in the response to the other reviewer, we believe that the high dimensionality should not have a significant effect on these results, even though we acknowledge that this effect would be interesting to study. Our justification have been added to the manuscript.

6. *Confusion matrix of the two best methods will be interesting.*

   We have computed the confusion matrices and they are indeed of interest as the two methods do not confuse the same classes. Though, due to page limit constraints, they can not be added to the manuscript.

7. *The cross-validation analyses confirmed that the value $T = 372$ ms for the support of the low-pass filter is optimal when applying late integration by majority voting, but shall it adapted per classes ?*

Although it is true that different values of T could be appropriate for different classes (that is, the different sound objects present could be of different scale), we do not expect results to change significantly. This is because the sounds are of a similar nature (ambient, non-directed) and so any distinct object shouldn't be much longer or shorter in duration in one class compared to another. In addition, $T = 372\,\text{ms}$ is at the longer end of the useful range of window sizes for which the scattering transform is useful, so it is unlikely that larger sizes would improve.

8. *A table with the 2016 DCASE private results may be added (or replace Tab III).*

   The private results have been added to the table.

9. *The bibliography is huge, it may be reduced by 10 %.*

   The bibliography section has been made shorter.

10. *Check typos in the references.*

    The references have been carefully curated.

11. *Bregman shall be cited in the discussion about binding of acoustic cues.*

    We sought for such a discussion in the manuscript and were unable to find one.