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

December 1, 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, alpha may appear in previous equations*
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*
3. *The class accuracy Table III for Home and Residential area are singular. Authors shall discuss that.*
4. *Complete legend fig 5.*

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.*
2. *The second paragraph of the abstract is vague.*
3. *No novel technique is presented*
4. *the whole paper architecture is poor*

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

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

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

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

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

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.*

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

12. *The function phi 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 phi is a scaling function.*

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

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

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

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".*

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

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

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

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

21. *Equations 9 and 10 are superfluous.* Done.

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.*

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

3. *Setting of Q1 shall be motivated.*

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

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.*

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

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 ?*

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

9. *The bibliography is huge, it may be reduced by 10 befor Bregman shall be cited in the discussion about binding of acoustic cues.*