Dear editor, dear reviewers,

Thank you for your helpful comments and suggestions. We have addressed your comments by reorganizing the outline of the paper, expanding our discussion of related work, and refining the less clear aspects of our methodology. With this letter, we resubmit the revised version of our manuscript. We hope that you will find these changes satisfactory.

Please find our responses to the reviewers' individual comments below.

Sincerely,
Vincent Lostanlen, Christian El-Hajj, Mathias Rossignol, Grégoire Lafay, Joakim Andén, and Mathieu Lagrange

REVIEWER 1

Invariance across pitch and dynamics: The authors claim that by first approximation timbre is invariant of pitch and dynamics. It should yet be noted that timbre invariance across pitch was shown to depend on musical training (Steele & Williams, 2006). Hence, the extent to which generalizations across pitch and dynamics are valid requires discussion, as well as the extent to which test participants were musically trained (although it is not clear to me whether any data on this were collected).

We have added a paragraph in the "Semi-supervised label propagation" subsection to discuss the joint effects of pitch and musical training in the modeling of auditory similarities between instrumental playing techniques. Specifically, we have cited the publications of Handel and Erickson (2001) ; Marozeau et al. (2003) ; and Steele and Williams (2006).

Furthermore, we have added a paragraph in the "Efficient annotation interface" subsection to explain why we did not collect personal information on the musical background of participants. That being said, since the participants were either students from the CNSMDP or persons that have subscribed to music or audio related mailing list, we can assume some sort of musical expertise.

That question of timbre invariance across pitch is a very important research question, that is somewhat taken for granted in our experimental protocol. We plan in future investigation to address this matter more carefully.

Role of instrument categories: Lemaitre et al. (2010, JEP) and Siedenburg et al. (2016, Frontiers) have shown that even qualitative ratings supposedly solely reflecting acoustic similarity are infiltrated by cues based on instrument family, resonator type, or type of excitation (i.e., technique). This could also be the case for the present sorting task in which case participants would only partially sort according to qualitative similarity. This should be critically discussed.

This point is well taken. However, our experimental protocol does not hypothesize that our annotation interface (Cyberlioz) only collects acoustical similarity, nor that it is entirely decorrelated from categorical similarity. In fact, our study confirms the findings of Siedenburg et al.: as shown in Figure 2 (inter-instrument similarity matrix), the consensus clustering arising from qualitative similarity judgments aligns well with instrument categories.
We have cited Siedenburg et al. (2016) in our "Related work" section to clarify this point.

Experimental methods: note that Giordano et al. (2011) demonstrated that free sorting of stimuli as a means to gather similarity data (such as used in the present manuscript) has considerable weaknesses in terms of reliability. These aspects require discussion in the manuscript. Furthermore, Elliott et al. (2013) collected direct dissimilarity ratings for 42 sounds (although not all participants rated all pairs of sounds). The data were analysed using the SMACOF algorithm for multidimensional scaling. Hence this is an efficient approach to get reduce the quadratic complexity of direct ratings and should be mentioned on p. 8 in the respective discussion of the experimental task.

We thank Reviewer 1 for these suggestions. We were not aware of the limitations of free sorting at the time of data collection. We have cited Giordano et al. (2011) and Elliott et al. (2013) in our "Efficient annotation interface" subsection.

Please better describe the experimental task. E.g. were all stimuli presented at once or in several blocks? What were the instructions to participants?

All the stimuli were presented at once. We have added a sentence in the "Efficient annotation interface" subsection to clarify this point. The subjects were asked to ‘cluster sounds into groups by assigning the same color to the corresponding dots according to how similar the sounds are.’ The subjects were allowed to move the dots on the plane, and their location was recorded, but it is not used to infer the perceptual similarity. No explicit guidelines were given to the subjects to substantify the notion of similarity, hopefully leading to an estimate of a measure of the ‘spontaneous’ similarity.

The manuscript claims that the system could be adapted to individual listeners but does not address or quantify inter-individual variability. How consistent were responses across participants?
To answer this question, we have added a new figure ("Inter-subject variability", now Figure 2) with the histogram of the number of clusters and the size of the clusters. We have discussed the results of this figure in the "Hypergraph partitioning" subsection.

Importantly, the structuring of the paper needs major improvement in order to be accessible for the reader. Specifically, several sections should be renamed, and the paper should be made more concise: - In the introduction, the subsection denoted as "Contributions" should better be called "Approach", because it mainly describes methods but not results. - The subsection called "Outline" can be fully removed, because

it does not contain any relevant information. - The subsection "Claim of originality" seems to be highly redundant with "Contributions". - The subsections "Semi-supervised label propagation" and "Evaluation metric" are part of the Methods, not the Results.

We thank Reviewer 1 for these suggestions. We have reorganized the paper and modified section titles accordingly.

The subsection "Visualization" should go into the Results section, if deemed relevant. Personally, I don't see the importance of this part of the study. The visuals could remain for illustrative purposes, but this does not need to be advertised as a main contribution.

For the sake of conciseness, we have removed the "Visualization" subsection from the present paper.

The whole Discussion section should be part of the Results section, because here most of the actual and interesting results are being presented. It would fit the more technical style of writing of a journal such as EURASIP if the authors would get rid of the questions in these titles (e.g., "Is metric learning necessary?" -> "The role of metric learning") - It seems reasonable to split the methods of data collection and classification modeling, the reader could be made more aware of this by providing more specific section titles, e.g., "Data collection" -> "Perceptual data collection" and "Methods" -> "Classification methods".

We have reorganized the paper and modified subsection titles accordingly.

There are several passages which simply repeat information from previous parts of the manuscript. While I don't think this is necessarily bad, I had the impression it was too much here, making the manuscript unnecessarily long.
We have removed repetitions in the paper whenever possible.

The subsections "Inter-instrument similarity" and "Inter-technique similarity" should be part of the Results section (because they present results). Even more importantly, it has not become clear to me what the implications of these analyses are - how exactly do the findings here fit into the overall picture of the paper? Or could they be placed to an appendix? Given the relatively broad dataset, I thought it would have been interesting to quantitatively compare inter-instrument similarities and inter-technique similarities, which could be an interesting addition or project for future research.

For the sake of conciseness, we have removed these subsections from the present paper. We agree with Reviewer 1 that these findings deserve further consideration as part of a future research.

The subsection "Timbre modeling in music cognition research" is heavily focussed on STRF approaches. Hence it should be called more specifically something like "STRFs for

timbre modeling". The present title is far to general; there would be much more to say about timbre modeling beyond STRFs than presently addressed in the manuscript, e.g. see Agus et al. (2012), Ogg et al. (2017), Siedenburg et al. (2019), and Caetano et al. (2019).

We thank Reviewer 1 for these references. We have modified the subsection title accordingly and have cited the recently published chapter of Caetano et al. (2019) for further detail on timbre modeling beyond STRF.

For better accessibility to non-musically adept readers, a short glossary on musical terms such as vibrato and tremolo could be added.

We have added a glossary of musical terms and abbreviations.

Importance of joint spectrotemporal modulations: P. 21, 2nd par: cite Patil et al (2013) here, who found similar pattern of results. But also see Siedenburg et al. (2019), who observed similar performance in instrument identification of a human and machine classifier using separable Gabor filterbank features.

We have cited and discussed the results of Patil et al. (2013) as well as Siedenburg et al. (2019).

Minor comments: [...]

We thank Reviewer 1 for their careful review. We have addressed all minor comments.

REVIEWER 2

My only reservation about this paper is that the authors do not discuss the difficulty, or lack thereof, of performing listening tests and collecting responses from several users for a large database on sounds.

We have revised the "Perceptual data collection" section so as to include more details on the tradeoffs of the free sorting task in comparison to the direct collection of pairwise similarity ratings, both in terms of efficiency and accuracy.

Minor errors/suggestions: [...]

We thank Reviewer 2 for their careful review. We have addressed all minor comments.

P9: Is there a reason for choosing a color pallette of 20? Would be interesting to see if users would use more or fewer colors based on the maximum number of colors allowed.

Although the effect of the maximum number of colors allowed is beyond the scope of our study, we have included a new figure (Figure 2) presenting a histogram of the number of clusters (i.e., colors) per participants. This figure shows that setting the maximum to 20 did not constrain too much the subjects in their choices.

Semi-supervised label propagation: is there a reference for this? Label propagation methods have a specific meaning in semi-supervised learning

This point is well taken. We have renamed the subsection "Semi-supervised label propagation" to "Extension to diverse pitches and dynamics" to avoid confusion.

Figure 1: Diagram can be improved. If an even higher-level view is shows at the beginning, that might be more helpful. It is difficult to understand the figure at first glance. If changing the figure is difficult, maybe instead of only referring to the intro, a concise description of a user's interaction with the proposed system can be provided in addition.

We have added a "Use case" subsection in the introduction to clarify this point.