For research article

Response to Reviewer 1 Comments

1. Summary

Thank you very much for taking the time to review this manuscript. We greatly appreciate your highly insightful and valuable comments. Please find the detailed responses below and the corresponding revisions in the re-submitted files.

2. Questions for General Evaluation	**Reviewer's Evaluation**	**Response and Revisions**
Does the introduction provide sufficient background and include all relevant references?	Yes	We have added a diagram of our method and its explanation.
Is the research design appropriate?	Must be improved	We have added further clarification regarding the motivation and effectiveness of the proposed method.
Are the methods adequately described?	Must be improved	Supplementary explanations have been added.
Are the results clearly presented?	Can be improved	Supplementary explanations have been added.
Are the conclusions supported by the results?	Must be improved	We have added further clarification regarding the significance of the results.

3. Point-by-point response to Comments and Suggestions for Authors

Comments 1: *First, a diagram outlining each step of the proposed method is necessary to visualize the integration of the data (and its modifications) with the proposed classifier.*

Response 1: Thank you for pointing this out. We have added a diagram of our method in Figure 1.

Comments 2: *The authors state that the "absence" class is not included in their classifier because it would be easy to integrate it with PIR sensors. This would mean integrating another system in addition to the microphones. Were experiments conducted considering this class with only the acoustic signals?*

Response 2: Yes, we conducted experiments to cluster activities solely from acoustic signals including "absence" activity. However, the performance was not satisfactory. The reason seems to be that codes corresponding to silence during activities may be common with codes

representing “absence”. Assuming that absence activity could be identified by other sensors, by removing codes appearing in “absence”, silence-related codes in activity can also be removed. As you pointed out, identifying “absence” activity would require another system including such as PIR motion sensors. We consider that the use of such a system can be a realistic solution, and thus we believe that we have demonstrated a practical and effective recipe for unsupervised activity recognition based on household sounds. We have added a note about “absence” activity to 4th paragraph in Section 6 and described the prospects for combining sensors with different modalities in 7th paragraph of Section 6.

Comments 3: *Further explanation is needed regarding the modifications made to the original data, specifically the discretization and beamforming:*

a) How does the residual vector quantizer work and the effects in the classifier?

c) The procedure: This residual is subsequently projected onto the closest entry in a second codebook, and a new residual is again computed. By repeating this process eight times, a sequence of eight codes representing the input embedding is obtained. Why is this done like that? It is not clear.

Response 3: Thank you for pointing out our lack of clarity. The residual vector quantizer (RVQ) is a function implemented within the neural audio codec, not a function implemented in the proposed method. RVQ is designed to control accuracy of encoding by the number of quantization stages. Increasing the number of quantization stages, that is, increasing the number of codes per frame allows the input frame to be encoded with higher precision. By default, a sequence of eight codes encodes each audio frame, but the proposed method uses only the first two codes as an acoustic word. This corresponds to performing an approximation that is even coarser than RVQ. While this coarse approximation reduces computational complexity, what this study has shown is that, unlike speech recognition, such a coarse approximation is sufficient to distinguish activities in activity recognition. We have added a description clarifying the meaning of RVQ and the design intent behind the acoustic words we adopted at the end of the 1st paragraph in Section 4.1.

Comments 4: *Further explanation is needed regarding the modifications made to the original data, specifically the discretization and beamforming:*

b) How the entire dataset is expanded, such that each original audio segment corresponds to three monaural audio recordings? How are you sure that the used software is really generating signals like real audio signals (bouncing in multiple directions)?

Response 4: Yes, each original 4 channel audio segment corresponds to three monaural audio recordings. Each monaural recording is computed using the classical delay-and-sum (DAS) beamforming. Specifically, after time delays are applied to signals of each microphone to compensate for arrival time differences from a particular direction, the delayed and aligned signals are then summed from all microphones. This process is expected to enhance direct sounds coming from the specified direction. In this study, by using three input streams in which each sound is enhanced for each predetermined direction, we aim to perform clustering that incorporates spatial information within household sounds. While the DAS beamforming may not achieve precise sound source separation and the signal can be distorted, we expect to improve clustering performance by leveraging the spatial information manifested as difference in the encoding of each stream depending on whether the sound is directional or omnidirectional. We have added

supplemental explanations about beamforming to Section 3.2 (from line 175 to line 178) and to the 3rd paragraph in Section 6.

4. Response to Comments on the Quality of English Language
Point 1: [No issues were noted.]

Response 1: We have corrected typographical errors in the abstract and improved the phrasing in the last section.

5. Additional clarifications

[Here, mention any other clarifications you would like to provide to the journal editor/reviewer.]