For research article

Response to Reviewer 2 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?	Can be improved	Additional explanations have been added.
Is the research design appropriate?	Can be improved	We have added further clarification regarding the motivation and effectiveness of the proposed method.
Are the methods adequately described?	Can be improved	Additional explanations have been added.
Are the results clearly presented?	Must be improved	Additional explanations have been added.
Are the conclusions supported by the results?	Must be improved	We have included additional explanations of conventional methods and added comparative experiments.

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

Comments 1: *The presence of other environmental sounds (TV, noise made by dropped objects, etc) are acknowledged but not explicitly modeled or mitigated in the experiments.*

Response 1: Thank you for pointing out that our explanation regarding how to mitigate the effects of noise was insufficient. When noise sources can be anticipated in advance, for example, the sound of a television, we assume that noise suppression techniques such as sound source separation could be applied, although this would depend on the overall cost. However, in the present study, we have not yet explored a thorough investigation of the signal processing performed prior to the neural audio codec (NAC), and we recognize this

as an issue for future work.

Regarding the noise reduction after NAC, codes corresponding to silent region during activities are noise appearing frequently. These codes are removed by removing codes appearing 'absence' activity, which is assumed to be detected by other sensors such as PIR motion sensors. We also expect that low-frequency codes derived from noise will be removed by the term frequency ratio.

Moreover, for codes originating from noise with a certain degree of overall frequency, we expect that the DMM model, which captures the burstiness of codes, to suppress the influence of noise-derived codes. This is because noise-derived codes may not exhibit burstiness compared with codes related to the activity of each segment. However, we have not yet been able to conduct a detailed analysis of this effect.

In response to your comment regarding noise-handling strategies, we have added a discussion of these issues as future work to the 3rd and 4th paragraph in Section 6.

Comments 2: *The authors claim that no explicit procedure for selection is required, but performance still depends on term frequency ratio thresholds and large preset values, which are manually chosen and not learned.*

Response 2: As you pointed out, the term frequency ratio is a parameter that must be set in advance, and it affects clustering performance. Our experiments indicate that this threshold should be set as large as possible within the range where computational stability is not compromised. Since computational stability can be analyzed and measured, we consider that determining the threshold is easier than determining the number of clusters. Nevertheless, in addition to this parameter, several other system-level design choices must be made in order to develop a practical system. We have added a description of the design choices that require further consideration to the 8th paragraph of Section 6.

Comments 3: *The approach depends on microphone array beamforming, which may be impractical or cost-prohibitive in real home deployments.*

Response 3: Our proposed method does not depend on a microphone array and can operate using a monaural microphone as input. However, employing microphone arrays enables the use of spatial information contained in household sounds and allows for the enhancement of sounds originating from specific directions. As you pointed out, microphone arrays are more expensive than other sensors such as PIR motion sensors, and their long-term operation requires a continuous power supply, which leads to greater installation constraints. Therefore, we also consider it impractical to deploy multiple microphone arrays. Nonetheless, compared with sensors that simply detect the location where an activity occurs, acoustic sensors have the advantage of being able to distinguish between different activities performed in the same location. Moreover, the cost of microphone array modules has been decreasing in recent years, for example, "reSpeaker XMOS XVF3800", 4-channel microphone-array module based on integrated DSP-enabled chips, can be obtained for approximately 55 USD. For these reasons, we consider it both practical and effective to use microphone arrays in a complementary manner with inexpensive and easy-to-use sensors, such as PIR motion sensors, deploying them only in locations where their advantages can be fully utilized. Your comment made us realize that it would be better to explain why the acoustic sensor is helpful, so we have added a sentence to the 4th paragraph of Section1 (from line 49 to line 52), and four sentences at the

beginning of 7th paragraph of Section 6.

Comments 4: *The experiments rely on the DCASE2018 dataset, which contains recordings from only one individual (young male). This severely limits the generalizability of the method, especially given the paper's target application to elderly living alone.*

Response 4: As you correctly pointed out, the dataset used in this study does not consist of the daily living sounds of older adults, who are the target population of this research. Furthermore, is contains data from only a single individual. Consequently, the effectiveness evaluation in this work is subject to certain limitations. However, based on our experience, variations in household sounds caused by age-related declines in physical function are relatively minor in the context of activity recognition, which contrasts to speech recognition. Instead, differences arising from individuals' personal belongings and lifestyle habits have a much greater impact. For this reason, together with scarcity of household sound dataset targeted the elderly, we employed the dataset. Even though the data were collected from only one young male, we consider it meaningful that our approach demonstrates the ability to distinguish activities from household sounds without any annotation of activities. We have added a note at the end of the 1st paragraph of Section 6.

Comments 5: *Comparisons are mostly restricted to K-Means and Spectral Clustering. More recent unsupervised or self-supervised audio representation learning methods must be considered.*

Response 5: We would appreciate any correction if our understanding is inaccurate, but, to our best knowledge, the study by Wang+ in 2020 [33] is the most promising results on unsupervised ASC. They use a custom neural autoencoder to extract embedding vectors from the bottleneck layer. For each audio segment, a feature vector is obtained by averaging embedding vectors of all frames. Then spectral clustering is applied to the set of feature vectors to obtain clusters. They also demonstrate that concatenating feature vectors obtained from different channels to incorporate spatial information further improves performance. Unfortunately, the dataset they used is not a dataset for the activity recognition, so it is not possible to directly compare their results with ours, nor could we use their code. Therefore, we attempted to apply their method on DCASE2018 Task5 dataset by using latest neural autoencoder pretrained on large data. Reflecting your comment, we have added a more detailed explanation of prior work [33] to the 4th paragraph of Section 2 and explicitly stated that we apply their method to the target dataset by using the latest neural autoencoder.

Your comment made us realize that it would be better to compare with existing clustering methods on acoustic words employed in acoustic topic model (ATM) [29]. Although ATM is not trained entirely through unsupervised learning, it can operate as such by assigning to each segment the most frequent cluster. Therefore, we conducted additional experiments using the Latent Dirichlet Allocation (LDA) model employed in [29]. As anticipated from our previous work [9], when burstiness is observed in the data, the DMM model exhibits better performance than the LDA model. We have added this result to Figure 8 and explained the result starting from 2nd paragraph of Section 5.1.

Comments 6: *Gibbs sampling for multi-stream DMMs may be computationally expensive, but runtime, convergence behavior, and scalability are not discussed.*

Response 6: Thank you for pointing out the important issue we failed to mention.

Although Gibbs sampling is guaranteed to converge to the target distribution in its stationary

distribution, we have not discussed its convergence properties. In fact, in our experiments, we did not perform any convergence diagnostics. Instead, when setting the number of iterations to 50, we simply confirmed by visual inspection that the parameters appeared stable. Convergence assessment is an important issue for future work.

We agree that the computational cost of the multi-stream DMM may not be lightweight enough for edge devices. However, its computational cost scales linearly with all prameters, such as the number of streams, the number of segments, the number of acoustic words in each segment and the number of clusters. Actual execution time was 393 seconds for clustering 13,531 audio segments, and 0.13 seconds for estimating a cluster of a 10-second audio segment on AMD Ryzen 9 5950X 2.2GHz CPU. These times do not include the time required for encoding, and the number of clusters was set to 50. Based on these facts, we consider that our clustering method is efficiently executable on the cloud.

We have added a description about time and space complexity immediately before the last paragraph of Section 4.5. We also add a note on the convergence behavior, runtime and scalability to the 5th and 6th paragraph of 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 Conclusion section.

5. Additional clarifications

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