

The authors bear all responsibility in case of violation rights. Upon acceptance, the dataset and code of benchmark models will be publicly released on GitHub under the CC-BY 4.0 license.

1 Motivation

For what purpose was the dataset created? The dataset was created for the purpose of creating sleep event monitoring systems on wearable devices.

Who created the dataset, and on behalf of which entity? Omitted.

Who funded the creation of the dataset? Omitted.

2 Composition

What do the instances that comprise the dataset represent? Each instance is an audio clip and imu data clip, as well as event category and interval.

How many instances are there in total? 420 hours audio data.

Does the dataset contain all possible instances or is it a sample of instances from a larger set? Contain all possible instances.

What data does each instance consist of? Each instance is an audio clip and imu data clip, as well as event category and interval.

Is there a label or target associated with each instance? Yes, audio and imu data are labeled with event category and event interval.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit? Yes.

Are there recommended data splits? Yes.

Are there any errors, sources of noise, or redundancies in the dataset? Yes, data are labeled manually and we discussed the possible error in Section 3.

Is the dataset self-contained, or does it link to or otherwise rely on external resources? The dataset is self-contained.

Does the dataset contain data that might be considered confidential? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset identify any subpopulations? No.

Is it possible to identify individuals, either directly or indirectly from the dataset? It may be possible. Audios may identify individual people, but the owners of the audios agreed for the dataset publishing, and we remove the conversations that might expose user's voice feature.

Does the data contain data that might be considered sensitive in any way? Yes, see above.

3 Collection process

How was the data associated with each instance acquired? Data collection details are provided in Section 3.1 of the paper.

What mechanisms or procedures were used to collect the data? See Section 3.1 of our paper.

If the dataset is a sample from a larger set, what was the sampling strategy? No, the whole dataset are collected from scratch.

Who was involved in the data collection process and how were they compensated? Data collection was done by the authors. Participants were paid 70 USD per night sleep, with the hourly wage estimated to be 10 USD. We collected 62 nights data in total and spent 4340 USD.

Over what timeframe was the data collected? The data was collected between June 2023 and September 2023.

Were any ethical review processes conducted? By the IRB of university.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources? Data was collected via the commercial hardwares and from recruiting participants.

Were the individuals in question notified about the data collection? N/A.

Did the individuals in question consent to the collection and use of their data? N/A.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects been conducted? No.

4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done? Data are manually annotated by annotaters with experience. All the data are not cleaned or preprocessed for the dataset.

Is the software that was used to preprocess/clean/label the data available? Audacity.

5 Uses

Has the dataset been used for any tasks already? Yes, it has been used for the experiments detailed in the paper.

Is there a repository that links to any or all papers or systems that use the dataset? No.

What (other) tasks could the dataset be used for? The dataset could be used for a variety of tasks including audio classification, audio model pretraining, sound event detection.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? Possibly – this varies with the application it is used for.
Are there tasks for which the dataset should not be used? Judgment should be used for tasks that may directly affect real people.

6 Distribution

Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created? Yes, it will be publicly available.

How will the dataset be distributed? Via GitHub and huggingface.

When will the dataset be distributed? The data is released on GitHub and huggingface, and will be finalized before the camera-ready version is submitted.

Will the dataset be distributed under a copyright or other intellectual property license, and/or under applicable terms of use? The dataset will be licensed under CC-BY 4.0.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The first author will maintain the dataset and it will be hosted on GitHub.

How can the owner/curator/manager of the dataset be contacted? They can be contacted via email.

Is there an erratum? An erratum will be included in the repository if necessary.

Will the dataset be updated? The dataset will be updated to correct any errors that are found.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances? Yes, all the collected data will be retained as long as the participants do not change their minds. If an participants require us to remove their content, it will no longer be accessible for the dataset.

Will older version of the dataset continue to be supported/hosted/maintained? No.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, they are free to fork the GitHub repository.