# Multi-Modal Dataset Across Exertion Levels: Capturing Post-Exercise Speech, Breathing, and Phonocardiogram

Jingping Nie[1], Yuang Fan[1], Minghui Zhao[1], Runxi Wan[2], Ziyi Xuan[1],
Matthias Preindl[1], and Xiaofan Jiang[1]
[1]Columbia University, [2]Tenafly High School
{jn2551, yf2676, mz2866, zx2420, matthias.preindl}@columbia.edu, wanrunxi838@gmail.com, jiang@ee.columbia.edu

## ABSTRACT

Cardio exercise elevates both heart rate and respiration rate, resulting in distinct physiological changes that affect speech patterns, pitch, breathing sounds, and heart sounds. These variations, which occur post-exercise, are influenced by factors such as exercise intensity and individual fitness levels. A comprehensive audio dataset is critically needed to capture post-exercise physiological changes, as existing datasets focus mainly on resting speech, breathing, and heart sounds, neglecting the dynamic shifts following physical exertion. Current datasets fail to capture unique post-exercise variations like speech disfluencies, altered breathing patterns, and variable heart sound intensities, limiting model generalizability to post-exercise conditions. To address this gap, we recruited 59 subjects from diverse backgrounds to engage in cardio exercise, specifically running, reaching varied exertion levels to produce a rich dataset. Our dataset includes 250 sessions totaling 143 minutes of structured reading, 47 minutes of spontaneous speech, 71 minutes of breathing sounds, and 62.5 minutes of phonocardiogram (PCG) recordings. We designed and deployed preliminary case studies to show that speech changes post-cardio could serve as an indicator of exertion level. We envision this dataset as a foundational resource for designing models in speech and cardiorespiratory monitoring that are resilient to the physiological shifts induced by exercise. This dataset could advance natural language processing (NLP) applications, mobile health, and wearable sensing technologies by enabling resilient and accurate physiological monitoring in real-world conditions.

## CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → *Health informatics*; • **Computing methodologies** → *Natural language processing*; • **Computer systems organization** → *Sensor networks*.

## KEYWORDS

Speech, Biosignals, Mobile health, Human-centered computing, Multimodal data


*SenSys '25, May 6–9, 2025, Irvine, CA, USA*

ACM ISBN 979-8-4007-1479-5/25/05
https://doi.org/10.1145/3715014.3722065

**ACM Reference Format:**
Jingping Nie[1], Yuang Fan[1], Minghui Zhao[1], Runxi Wan[2], Ziyi Xuan[1],, Matthias Preindl[1], and Xiaofan Jiang[1]. 2025. Multi-Modal Dataset Across Exertion Levels: Capturing Post-Exercise Speech, Breathing, and Phonocardiogram. In *The 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3715014.3722065

## 1 INTRODUCTION AND MOTIVATION

With the rise of artificial intelligence of things and ubiquitous smart mobile platforms, acoustic-based methods have gained interest in vital sign monitoring and fitness tracking, offering a non-invasive and unobtrusive alternative for health monitoring. Cardio exercise, known for its cardiovascular benefits, induces physiological changes affecting vital signs and the acoustic features of speech, breathing, and heart sounds [7]. Post-exercise breathing often shows distinct pauses and micro-breaths, with potential variations like exercise-induced wheeze or asthma [5, 41]. Since speech relies on breath coordination, altered respiration post-exercise may lead to disfluency, characterized by interruptions in the flow of speech, such as pauses, repetitions, or hesitations [59]. Heart sounds, or phonocardiograms (PCGs), also shift with increased cardiovascular workload, such as elevated heart rate and intensified first (S1) and second (S2) heart sounds [15].

Beyond fitness and health monitoring, this work intersects with advancing fields of natural language processing (NLP) and automatic speech recognition (ASR), which are transforming human-technology interactions. Leveraging microphones in mobile devices for acoustic-based health monitoring offers a unique dual advantage: enabling multi-task applications that benefit both health and speech-related functionalities [22, 25, 32, 46, 68]. Current ASR and voice-assistant systems often underperform for users who stutter, are non-native speakers, have accents, or have speech disabilities [43, 55]. ASR systems also struggle to accurately recognize children's speech due to differences in speech patterns [2]. Integrating speech analysis with disfluency and breathing pattern detection enhances ASR systems to accommodate atypical breathing and semantic breaks. This approach also aids in assessing and treating stuttering and conditions like Parkinson's disease [41, 49, 58, 69].

However, existing cardiorespiratory audio datasets and speech breathing/disfluency datasets are limited in capturing the unique acoustic variations that occur post-exercise. These datasets predominantly focus on resting speech, breathing, and heart sounds, overlooking the dynamic physiological changes that follow physical exertion [1, 14, 17, 18, 21, 30, 35, 47, 53]. These datasets lack the breadth of acoustic data reflecting speech disfluencies, altered breathing patterns, and the varying intensities of heart sounds that

emerge after cardio exercise, making it challenging for current models to generalize to post-exercise conditions [46].

To address this gap, we created a novel dataset by recruiting 59 participants from diverse demographic backgrounds and fitness levels to engage in running as a means to achieve various exertion levels while recording acoustic data. Although running was chosen for this study, its sole purpose was to modulate exertion levels after cardio exercises, as research indicates that different cardio exercises can be adjusted to induce similar exertion levels by modifying intensity and duration [33, 63]. Our dataset comprises over 250 sessions, including 143 minutes of structured reading (read paragraphs out loud), 47 minutes of spontaneous speech, 71 minutes of breathing sounds, and 62.5 minutes of phonocardiogram (PCG) data. Additionally, the dataset contains non-sensitive background information, such as age, gender, weight, and exercise and running experience, allowing for more nuanced analyses. The details of this proposed dataset are introduced in Section 3. We conducted a case study with this dataset, which leverages post-exercise speech patterns to classify exertion, obtaining a 5-fold cross-validation accuracy of 73.02% on structured reading and 81.02% on spontaneous speech, offering a promising avenue for personal health monitoring. This dataset provides a foundation for (*i*) developing acoustic models resilient to variations in post-exercise physiological exertion levels, (*ii*) advancing embedded AI and sensing applications in fitness tracking, health monitoring, and speech technology, and (*iii*) expanding acoustic-based monitoring in mobile health and wearable devices while enhancing the adaptability of voice-based interfaces. The dataset is available at https://github.com/Columbia-ICSL/data_after_cardio.

## 2 BACKGROUND AND RELATED WORK

**Cardiorespiratory Audio Datasets and Mobile Sensing Applications:** Datasets in the literature containing cardiorespiratory audio, such as Audio Set [17], FSD50K [14], and FluSense [1], are typically annotated for audio event classification and are often short in duration and lack in a post-exercise context. A few datasets provide more detailed annotations specifically for respiratory and cardiac sounds. For example, the dataset from [53] includes timestamped annotations of respiratory cycles and labels for wheezing and crackles. Existing research utilizing this dataset primarily focuses on automatic lung sound classification [16, 36, 39, 60]. More specialized datasets for heart sounds include the EPHNOGRAM dataset, which contains simultaneous electrocardiogram (ECG) and phonocardiogram (PCG) recordings, and the CirCor DigiScope Phonocardiogram dataset, which provides PCG recordings with detailed timestamped segmentations of S1 and S2 heart sounds, along with heart murmur annotations [30, 47]. Research on these datasets has led to advancements in PCG segmentation for detecting S1 and S2 sounds using deep recurrent neural networks (RNNs), contributing to improved heart rate estimation and abnormal heart sound detection [25, 28, 44, 46]. However, these datasets primarily capture heart and respiratory sounds at rest and do not account for the variations introduced by cardio exercise.

In addition to datasets, there have been several efforts to enable cardiorespiratory-related applications using mobile or wearable devices. Smartphones and headphones have been employed in a passive acoustic sensing system designed to detect rope-jumping activities and breathing patterns [26]. Ren et al. utilized smartphone-captured breathing sound after cardio exercise to monitor exercise intensity [52]. BreathPro was designed to monitor breathing modes during running [27]. RunBuddy leverages smartphones and Bluetooth headsets to monitor running rhythm and estimate local respiratory coefficient (LRC) [22]. A multi-task learning model was proposed to estimate respiratory rate from breath audio obtained through wearable microphones after exercise [32]. However, there is a lack of comprehensive datasets that capture the acoustic dynamics of cardio-respiratory sounds immediately following cardio exercise across varying exertion levels.

**Speech Breathing and Disfluency and Downstream Applications:** Cardio exercise may change speech breathing and introduce conversational disfluencies, including repetitions, restarts, and corrections. Previous research has explored machine learning approaches to estimate respiration rates from speech, such as using close-talking microphone recordings from subjects at rest [41]. However, while there exist datasets (such as Disfl-QA [21], Sep-28k [35], FluencyBank [51], and FluencyBank Timestamped [56]) that capture disfluent speech in various contexts, including speech recordings from individuals who stutter from children and adults or non-native speakers, as well as from podcasts, they primarily focus on speech produced in resting conditions [21, 35, 51, 56].

Recent advancements in automatic speech recognition (ASR) for stuttered or disfluent speech aim to enhance user experiences and expand applications for individuals who stutter, have speech-related conditions, speak with accents, or are non-native speakers [23, 42, 43]. Shonibare et al. introduced a method called 'Detect and Pass,' a context-aware classifier designed to improve ASR accessibility for individuals who stutter [58]. Additionally, multimodal architectures have been proposed to enhance disfluency detection accuracy compared to unimodal approaches [55]. Speech disfluency is also being utilized as a feature for disease assessment. For instance, PDAssess, a system employing free-speech analysis, provides a four-stage assessment of Parkinson's disease by analyzing disfluency and other speech characteristics [69]. Additionally, disfluencies in read speech have been shown to effectively predict cognitive impairment in individuals with Parkinson's disease [54]. However, despite these advancements, current datasets and methods primarily address disfluency in contexts without physical exertion, overlooking the distinct speech and breathing patterns that appear after cardio exercises.

## 3 EXPERIMENT SETUP AND DATASET

We design a portable and reliable data collection setup and experiment procedure, collaborating with a licensed running coach to ensure the safety of the participants. This research received approval from the Institutional Review Board (IRB). We recruit 59 adult subjects from diverse demographic distributions. Informed consent was obtained from all participants before their involvement in the study. Participants were fully informed about the nature of the data being collected, the purpose of the study, and their right to withdraw at any time. Non-personal-identifiable demographics and fitness level background information are collected.

**Experiment Procedure:** As illustrated in Figure 1, our experiment includes three main steps. First, in *Step 1*, each participant

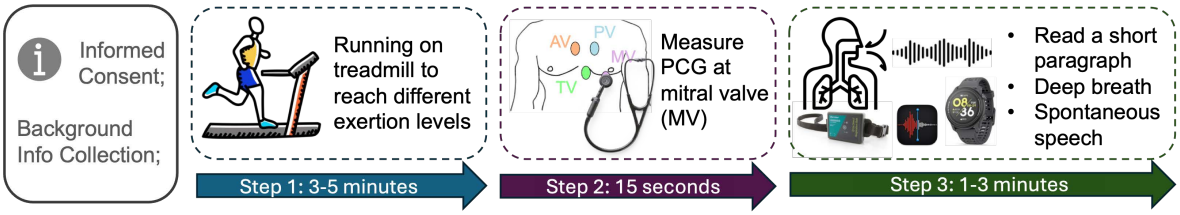


**Figure 1: The 3-Step experiment procedure. The proposed dataset includes: (*i*) the PCG recordings in *Step 2*; (*ii*) the reading, spontaneous speech, and deep breath sounds in *Step 3*.**

**Table 1: Exertion level definition of the modified five-level Borg RPE.**

| Exertion Level | Description |
|---|---|
| 1-*Very light* | Anything other than complete rest. |
| 2-*Light* | Can maintain for hours, easy to breathe and carry on a conversation. |
| 3-*Moderate* | Can exercise for long periods, able to talk, and hold short conversations. |
| 4-*Vigorous* | On the verge of becoming uncomfortable, short of breath, can speak a sentence. |
| 5-*Max effort* | Feels impossible to continue, completely out of breath, unable to talk. |

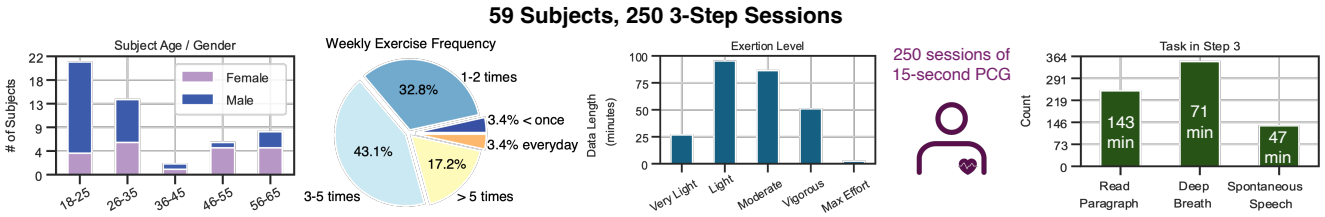


**Figure 2: Overview of the dataset to be open-sourced, comprising 250 3-Step experiment sessions from 59 subjects.**

completed several running sessions at constant speeds of 5, 6, 7, 8, 9, and/or 10 miles per hour (*mph*) for 3-5 minutes, based on their running proficiency, to achieve varying exertion levels. As shown in Table 1, to measure perceived exertion, we used a modified five-level Borg RPE (Rating of Perceived Exertion) scale [4], which is a well-established and widely-adopted method for assessing exercise exertion levels [32, 57, 66, 70]. Participants were informed that they do not require higher physical exertion than the "moderate" range. After each session, they reported their perceived exertion level. A certified running coach supervised the experiment to ensure participant safety and validate the accuracy of self-reported exertion levels.

Immediately following each running session, in *Step 2*, we recorded a 15-second phonocardiogram (PCG) at the mitral valve (MV) auscultation location—the heart's apex—using a 3M Littmann CORE Digital Stethoscope [24]. The MV location was selected for its clear access to mitral valve sounds, especially S1 (the first heart sound) and any murmurs associated with the mitral valve.

Finally, immediately after *Step 2*, each participant proceeded to *Step 3*, where they were asked to: (*i*) read a paragraph aloud provided by the experimenter, (*ii*) spontaneously describe their feelings post-run or share thoughts about their day while speaking alone, and (*iii*) take a few deep breaths, either before or after the reading and spontaneous speech. Participants wore a Coros Pace 3 sports watch [9] and a Vernier Go Direct strain-gauge chest-belt sensor [62] as references. Heart rate data was collected with the Coros Pace 3 (at 1 $Hz$), and chest expansion/contraction (indicating inhalation and exhalation) was recorded by the Vernier Go Direct respiration belt (at 20 $Hz$). Reading, spontaneous speech, and deep breathing sounds were recorded using the Voice Memos app on iPhones [29].

**Dataset:** Unix timestamps were used to synchronize the data sources: (*i*) heart rate from a sports watch, (*ii*) chest-belt pressure measurements (in Newtons), and (*iii*) audio capturing speech and respiration. Figure 2 provides an overview of the proposed dataset, which includes 250 3-Step experimental sessions from 59 subjects from diverse backgrounds and fitness levels. In particular, subjects reached various exertion levels after *Step 1* in the 250 sessions. There are 250 15-second PCG recordings collected in *Step 2*. Note that in *Step 3*, some subjects were unsure of what to say or spoke while laughing. We removed these invalid spontaneous speech clips. As such, there are 143 minutes of structured reading from 250 recordings, 47 minutes of spontaneous speech from 134 recordings, and 71 minutes of deep breathing sound from 347 recordings.

Figure 3 and Figure 4 illustrate: (*i*) the phonocardiograms (PCGs) recorded using a digital stethoscope, and (*ii*) spectrogram, heart rate, and chest-belt pressure measurements when reading an identical paragraph immediately after an example subject completed 5-minute treadmill running sessions at 6 *mph* and 10 *mph*. This subject reported exertion levels of 2 (light) and 4 (Vigorous) to the 6-mph and 10-mph sessions, respectively. As shown in Figure 3, the PCG from the 10-mph session shows significantly higher amplitude and heartbeat frequency, indicating increased cardiovascular activity. In addition, when reading the same paragraph, as illustrated in Figure 4, the subject had more frequent breathing pauses, resulting in a longer time to complete the reading task after the 10-mph session. Micro-breathing was observed during the reading after both

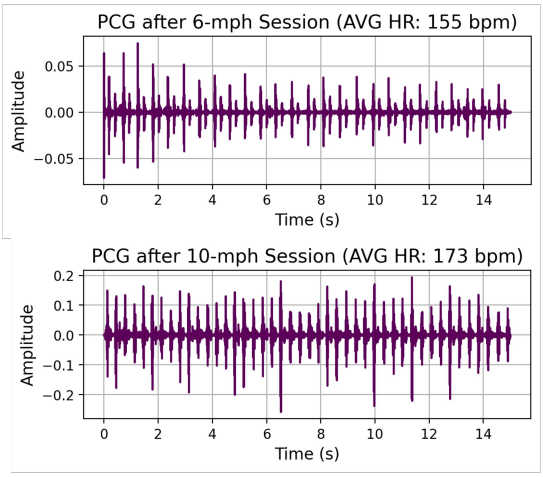


**Figure 3: Comparison of PCG recordings after a subject completed running sessions at 6 *mph* and 10 *mph* reveals variations in amplitude and interbeat intervals during the cooldown period.**

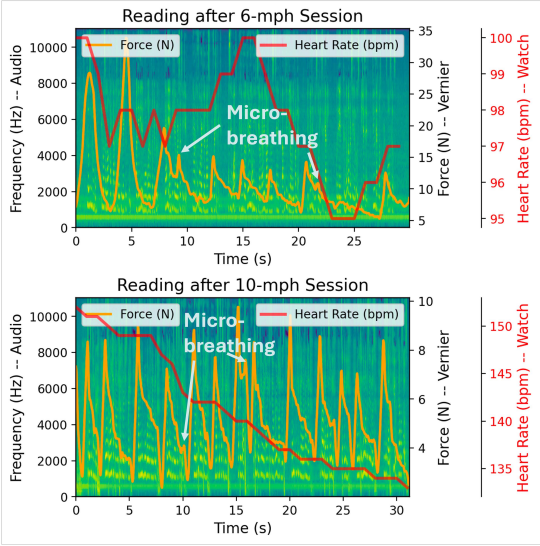


**Figure 4: Comparison for audio, breathing, and heart rate after a subject completed running sessions at 6 *mph* and 10 *mph*.**

6- and 10-mph sessions when the subject attempted to complete semantically connected sentences/phrases.

## 4 EXERTION LEVEL CLASSIFICATION CASE STUDY

To highlight the dataset's relevance for applications in fitness tracking and vital sign monitoring, we conducted a preliminary case study to benchmark popular classification models, demonstrating that post-exercise speech characteristics (audio recordings in *Step 3*) can indicate exertion levels.

**Features, Labels, and Model Training:** To provide a preliminary, interpretable analysis of exertion level classification, we grouped exertion levels into two categories: low (levels 1 and 2) and high (levels 3, 4, and 5), as self-reported RPE scores may contain personal biases. We investigated three features: (*i*) widely used acoustic

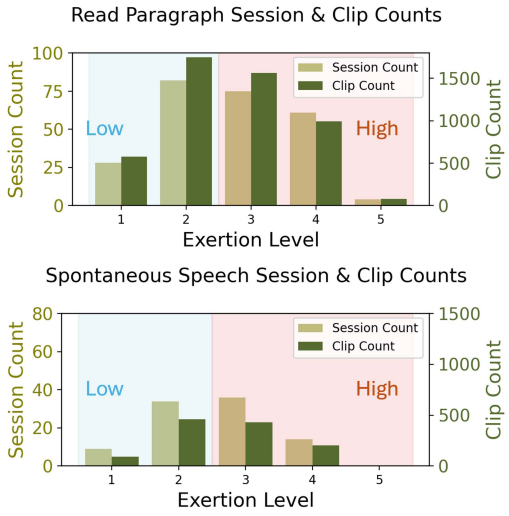


**Figure 5: Number of sessions and number of clips for structured paragraph reading and spontaneous speech.**

features, Mel Filterbanks (MFBs) and Mel-Frequency Cepstral Coefficients (MFCCs); and (*ii*) the layer 4 representations generated from a pre-trained Wav2Vec2-base model (W2V2 Emb-4), which have been shown to be informative for bioacoustic and speech breathing tasks in [32, 41, 46, 50]. We empirically segmented all audio recordings from the read paragraph and spontaneous speech tasks in *Step 3* into 15-second clips with a 1-second stride, resulting in 4,958 clips for the read paragraph task and 1,267 clips for spontaneous speech. The distribution of clips across exertion levels is shown in Figure 5.

With these features, we evaluated three neural network architectures: (*i*) a 4-layer multi-layer perceptron (`MLP`) model, (*ii*) a modified 2-dimensional convolutional neural network (`2D CNN`) model as proposed in [46], and (*iii*) a modified 1-dimensional convolutional neural network long short-term memory (`1D CNN-LSTM`) model from [41]. For each feature-model combination, we used recordings from the two speech tasks in *Step 3* (read a paragraph, spontaneous speech clips, and both) to perform 5-fold cross-validation with the segmented 15-second audio clips (randomly split on the session level) and evaluate how each speech type contributes to classification with the results shown in Table 2.

**Case Study Results:** The `1D CNN-LSTM` model consistently achieves high cross-validation accuracy across all speech conditions among the three models, with optimal results when using W2V2 Emb-4 (0.7302, 0.8102, and 0.7580 for Reading, Spontaneous Speech, and combined conditions, respectively), suggesting the `1D CNN-LSTM` effectively captures speech breathing characteristics. Overall, W2V2 Emb-4 contains the most information related to exertion level in speech breathing, followed by MFCC and then MFB. Additionally, we note that, across all model-feature combinations, classifying exertion level using only spontaneous speech clips outperforms using read paragraph clips or a combination of both, with slightly higher variability likely due to the smaller data size for spontaneous speech. This is likely because, during spontaneous speech, people adjust their breathing and speech tempo more naturally, while structured paragraph reading may lead individuals to rush through the text, distorting their natural speech breathing patterns

**Table 2: 5-Fold CV results of the combinations of model structures, features, and reading/spontaneous speech.**

| | | 5-Fold Average Cross-Validation (CV) Accuracy (mean ± std) | | |
|---|---|---|---|---|
| **Model** | **Feature (dim)** | **Reading** | **Spontaneous Speech** | **Reading & Spontaneous Speech** |
| MLP | MFB (40, 1292) | 0.5082 ± 0.05 | 0.5854 ± 0.03 | 0.5351 ± 0.04 |
| | MFCC (40, 1292) | 0.5115 ± 0.05 | 0.5651 ± 0.03 | 0.5364 ± 0.03 |
| | W2V2 Emb-4 (749, 768) | 0.6348 ± 0.03 | 0.7215 ± 0.07 | 0.6776 ± 0.02 |
| 2D CNN | MFB (40, 1292) | 0.6255 ± 0.05 | 0.7271 ± 0.05 | 0.6340 ± 0.03 |
| | MFCC (40, 1292) | 0.6752 ± 0.05 | 0.7329 ± 0.05 | 0.7061 ± 0.04 |
| | W2V2 Emb-4 (749, 768) | 0.212 ± 0.03 | 0.7687 ± 0.02 | 0.7040 ± 0.03 |
| 1D CNN-LSTM | MFB (40, 1292) | 0.6377 ± 0.07 | 0.7098 ± 0.06 | 0.6175 ± 0.03 |
| | MFCC (40, 1292) | 0.6725 ± 0.02 | 0.7469 ± 0.13 | 0.7173 ± 0.05 |
| | W2V2 Emb-4 (749, 768) | **0.7302 ± 0.05** | **0.8102 ± 0.04** | **0.7580 ± 0.03** |

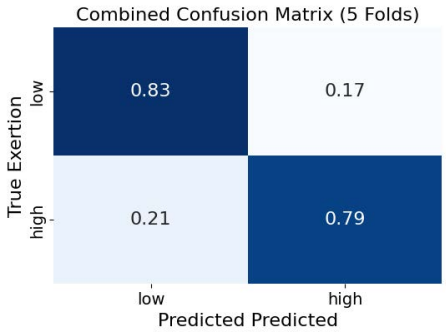


**Figure 6: Confusion matrix combining the results of 5-fold cross-validation for the 1D CNN-LSTM model using W2V2 Emb-4 features on spontaneous speech data.**

and acoustic features that may be highly correlated with exertion levels.

**Example Error Analysis:** We further looked into the 5-fold cross-validation results of the best-performing model (1D CNN-LSTM with W2V2 Emb-4 on spontaneous speech). The combined confusion matrix in Figure 6 reveals a slight tendency toward underestimation, where the model predicts low exertion when the true label is high. This may be caused by personal biases in perceived exertion and variations in individual recovery rates. Additionally, we observed that predictions remained consistent across 15-second window clips from the same recording, suggesting that exertion-related characteristics are consistently present in post-exercise speech.

## 5 FUTURE DIRECTIONS AND ENABLED APPLICATIONS

In this work, we recruited 59 subjects and created a comprehensive dataset that includes phonocardiograms, structured paragraph readings, spontaneous speech, and breathing (deep breaths) recorded after achieving various physiological exertion levels through cardio exercise and subject background information. Below, we outline future directions and potential applications enabled by this dataset:

**Exercise Physiology and Personalized Health Applications:** Beyond the preliminary case studies for exercise exertion monitoring in Section 4, this dataset can support more profound studies on recovery patterns for fitness and health monitoring. Integrating chest belt pressure, heart rate, and PCG signals can inform personalized training recommendations and detect abnormal recovery patterns indicative of health issues [38, 65]. Additionally, with access to subject background information, researchers can examine how factors such as age, fitness level, and exercise habits influence recovery rates and exertion levels, supporting personalized, real-time interventions in mobile health systems for preventive care and well-being management.

**Wearable and Mutimodal Sensing Technology:** The integration of speech breathing audio, chest-belt pressure measurements, and heart rate data in this dataset provides a foundation for advanced sensor fusion techniques in wellness tracking via personal mobile and wearable devices. This dataset offers valuable insights for designing and calibrating multimodal wearable systems capable of seamlessly combining respiratory, acoustic, and cardiovascular data to deliver continuous and unobtrusive health monitoring [10, 13, 20, 34, 45]. It can drive the development of innovative multimodal algorithms for virtual fitness assistants, allowing them to adapt recommendations based on real-time physiological responses and further extend the applications of embedded systems and wearable technology for personal self-care and proactive health monitoring [61].

**Machine Learning for Acoustic and Time-series Cardiorespiratory Signals:** As mentioned in Section 2, existing cardiorespiratory audio datasets [1, 14, 17, 47, 53, 67] primarily capture bioacoustic data at rest. While the proposed dataset is valuable as a standalone resource, its utility is further amplified when integrated with existing datasets, complementing them to enhance the development of machine learning and foundation models (such as CLAP [11], AST [19], and HeAR [3]), enabling more robust applications like audio event classification, disease detection, emotion classification, and health monitoring [12, 40], as well as supporting telemedicine by providing audio biomarkers insights to enrich the telemonitoring of patients [6]. For example, this dataset aids in distinguishing between pathological and exercise-induced cardiorespiratory conditions.

With data encompassing multiple time-series biosignals, bioacoustics, and structured/unstructured speech, this dataset can advance machine learning models that leverage multimodal inputs [31]. The dataset could also support the development of transfer learning

models to adapt across different exercise types, environments, and population demographics as well as multimodal time-series sensing, further bridging the gap between AI-driven health monitoring and practical deployment in mobile and embedded systems for everyday pervasive monioring [37].

**Speech and Language Analysis under Physical Exertion and Speech Breathing/Disfluency Analysis:** The proposed dataset can serve as a valuable addition to existing speech and disfluency datasets [8, 21, 35, 48, 51, 56], addressing a gap in contexts that involve physical exertion. By capturing speech, breathing, and disfluencies at different exertion levels, this dataset enables research into how physical strain influences speech production, breathing patterns, and disfluency types (e.g., pauses, repetitions, restarts), and further enhances applications in speech accessibility support, disease assessment (e.g., respiratory issues, neurological conditions, and speech disabilities), and emotion classification [23, 54, 69]. Additionally, popular voice-command systems often struggle with users who stutter or are physically active [64]. This dataset can help develop more robust ASR models and disfluency detection systems that account for physiological changes, making voice-command interfaces more accessible and effective in real-world settings. Furthermore, this dataset can advance real-time, on-device speech and breathing analysis in wearable and mobile systems and contribute to more adaptive human-computer interaction in dynamic environments for physically active user.

## 6 CONCLUSION

In this work, we present a comprehensive dataset that captures variations in speech, breathing, and heart sounds after cardio exercise at different exertion levels, addressing a critical gap in resources for analyzing physiological changes post-exercise. Our dataset includes 250 sessions from 59 subjects of diverse backgrounds and fitness levels, with 143 minutes of structured reading, 47 minutes of spontaneous speech, 71 minutes of breathing sounds, and 62.5 minutes of PCG data, spanning multiple modalities and exertion levels. It also incorporates background information such as age, gender, weight, and exercise experience, enabling in-depth analysis of physiological responses. Our preliminary case study demonstrates the potential of post-exercise speech features to accurately classify exertion levels, setting the stage for exertion tracking and extensive health monitoring applications. Future work can be built on these insights, developing adaptive wearable health systems, enhancing automatic speech recognition under physical strain, and supporting research in cardiorespiratory health and speech disfluency. Additionally, this dataset can contribute to advancing embedded AI and sensing applications, enabling more efficient real-time physiological monitoring and robust multimodal health analytics.

## ACKNOWLEDGEMENTS

This research was partially supported by the National Science Foundation under Grant Number 2133516 and Apple Inc. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Columbia University, NSF, or the U.S. Government or any of its agencies.

## REFERENCES

[1] Forsad Al Hossain, Andrew A Lover, George A Corey, Nicholas G Reich, and Tauhidur Rahman. 2020. FluSense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4, 1 (2020), 1–28.

[2] Sadeen Alharbi, Madina Hasan, Anthony JH Simons, Shelagh Brumfitt, and Phil Green. 2018. A lightly supervised approach to detect stuttering in children's speech. In *Proceedings of Interspeech 2018*. ISCA, 3433–3437.

[3] Sebastien Baur, Zaid Nabulsi, Wei-Hung Weng, Jake Garrison, Louis Blankemeier, Sam Fishman, Christina Chen, Sujay Kakarmath, Minyoi Maimbolwa, Nsala Sanjase, et al. 2024. HeAR–Health Acoustic Representations. *arXiv preprint arXiv:2403.02522* (2024).

[4] Gunnar Borg. 1998. *Borg's perceived exertion and pain scales*. Human Kinetics.

[5] British Journal of Sports Medicine. 2018. Your Patient Has an 'Exercise Associated Wheeze'... It Might Not Be Asthma! https://blogs.bmj.com/bjsm/2018/04/09/your-patient-has-an-exercise-associated-wheeze-it-might-not-be-asthma/. Accessed: 2024-11-09.

[6] Antonio Celesti, Marco Dell'Acqua, Giovanni Lonia, Davide Ciraolo, Fabrizio Celesti, Maria Fazio, Massimo Villari, Mirjam Bonanno, and Rocco Salvatore Calabrò. 2024. Leveraging Audio Biomarkers for Enriching the Tele-Monitoring of Patients. In *2024 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 1–6.

[7] Tao Chen, Yongjie Yang, Xiaoran Fan, Xiuzhen Guo, Jie Xiong, and Longfei Shangguan. 2024. Exploring the Feasibility of Remote Cardiac Auscultation Using Earphones. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 357–372.

[8] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5903–5917.

[9] Inc. COROS Wearables. 2023. COROS PACE 3 GPS Sport Watch. https://us.coros.com/pace3. Accessed: 2024-11-09.

[10] Han Ding, Longfei Shangguan, Zheng Yang, Jinsong Han, Zimu Zhou, Panlong Yang, Wei Xi, and Jizhong Zhao. 2015. FEMO: A platform for free-weight exercise monitoring with RFIDs. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 141–154.

[11] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[12] Zongyao Feng, Konstantin Markov, Junpei Saito, and Tomoko Matsui. 2024. Neural Cough Counter: A Novel Deep Learning Approach for Cough Detection and Monitoring. *IEEE Access* (2024).

[13] Glenn J Fernandes, Jiayi Zheng, Mahdi Pedram, Christopher Romano, Farzad Shahabi, Blaine Rothrock, Thomas Cohen, Helen Zhu, Tanmeet S Butani, Josiah Hester, et al. 2024. HabitSense: A Privacy-Aware, AI-Enhanced Multimodal Wearable Platform for mHealth Applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–48.

[14] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), 829–852.

[15] Kyle Fulghum and Bradford G Hill. 2018. Metabolic mechanisms of exercise-induced cardiac remodeling. *Frontiers in cardiovascular medicine* 5 (2018), 127.

[16] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. 2021. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 527–530.

[17] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.

[18] Rong Gong, Hongfei Xue, Lezhi Wang, Xin Xu, Qisheng Li, Lei Xie, Hui Bu, Shaomei Wu, Jiaming Zhou, Yong Qin, et al. 2024. AS-70: A Mandarin stuttered speech dataset for automatic speech recognition and stuttering event detection. *arXiv preprint arXiv:2406.07256* (2024).

[19] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[20] Xiaonan Guo, Jian Liu, Cong Shi, Hongbo Liu, Yingying Chen, and Mooi Choo Chuah. 2018. Device-free personalized fitness assistant using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.

[21] Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-QA: A Benchmark Dataset for Understanding Disfluencies in Question Answering. In *Findings of ACL*.

[22] Tian Hao, Guoliang Xing, and Gang Zhou. 2015. RunBuddy: a smartphone system for running rhythm monitoring. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 133–144.

[23] Mark Hasegawa-Johnson, Xiuwen Zheng, Heejin Kim, Clarion Mendes, Meg Dickinson, Erik Hege, Chris Zwilling, Marie Moore Channell, Laura Mattie, Heather Hodges, et al. 2024. Community-supported shared infrastructure in support of speech accessibility. *Journal of Speech, Language, and Hearing Research* (2024), 1–14.

[24] Eko Health. n.d.. 3M™ Littmann® CORE Digital Stethoscope. https://www.ekohealth.com/products/3m-littmann-core-digital-stethoscope?variant=39307014209632. Accessed: 2024-11-09.

[25] Kaiyuan Hou, Stephen Xia, Emily Bejerano, Junyi Wu, and Xiaofan Jiang. 2023. ARSteth: Enabling Home Self-Screening with AR-Assisted Intelligent Stethoscopes. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*. 205–218.

[26] Xiaowen Hou and Chao Liu. 2022. Rope Jumping Strength Monitoring on Smart Devices via Passive Acoustic Sensing. *Sensors* 22, 24 (2022), 9739.

[27] Changshuo Hu, Thivya Kandappu, Yang Liu, Cecilia Mascolo, and Dong Ma. 2024. BreathPro: Monitoring Breathing Mode during Running with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–25.

[28] Ahmed Imtiaz Humayun, Shabnam Ghaffarzadegan, Zhe Feng, and Taufiq Hasan. 2018. Learning front-end filter-bank parameters using convolutional neural networks for abnormal heart sound detection. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1408–1411.

[29] Apple Inc. 2024. Voice Memos. https://apps.apple.com/us/app/voice-memos/id1069512134. Accessed: 2024-11-09.

[30] Arsalan Kazemnejad, Sajjad Karimi, Peiman Gordany, Gari D Clifford, and Reza Sameni. 2024. An open-access simultaneous electrocardiogram and phonocardiogram database. *Physiological Measurement* 45, 5 (2024), 055005.

[31] Sein Kim, Namkyeong Lee, Junseok Lee, Dongmin Hyun, and Chanyoung Park. 2023. Heterogeneous graph learning for multi-modal medical data analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5141–5150.

[32] Agni Kumar, Vikramjit Mitra, Carolyn Oliver, Adeeti Ullal, Matt Biddulph, and Irida Mance. 2021. Estimating respiratory rate from breath audio obtained through wearable microphones. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 7310–7315.

[33] Anita Kumari, Swati Sinha, Amita Kumari, Anup Kumar D Dhanvijay, Sanjeet Kumar Singh, and Himel Mondal. 2023. Comparison of Cardiovascular Response to Lower Body and Whole Body Exercise Among Sedentary Young Adults. *Cureus* 15, 9 (2023).

[34] Guohao Lan, Tim Scargill, and Maria Gorlatova. 2022. Eyesyn: Psychology-inspired eye movement synthesis for gaze-based activity recognition. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 233–246.

[35] Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P Bigham. 2021. Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6798–6802.

[36] Shih-Hong Li, Bor-Shing Lin, Chen-Han Tsai, Cheng-Ta Yang, and Bor-Shyh Lin. 2017. Design of wearable breathing sound monitoring system for real-time wheeze detection. *Sensors* 17, 1 (2017), 171.

[37] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. 2024. FOCAL: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. *Advances in Neural Information Processing Systems* 36 (2024).

[38] Xinxin Ma, Xinhua Su, Huanmin Ge, and Yuru Chen. 2024. PCG-based exercise fatigue detection method using multi-scale feature fusion model. *Computer Methods in Biomechanics and Biomedical Engineering* (2024), 1–14.

[39] Yi Ma, Xinzi Xu, and Yongfu Li. 2020. LungRN+ NL: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation.. In *Interspeech*. 2902–2906.

[40] George Mathew, Daniel Barbosa, John Prince, and Subramaniam Venkatraman. 2024. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. *npj Cardiovascular Health* 1, 1 (2024), 25.

[41] Vikramjit Mitra, Anirban Chatterjee, Ke Zhai, Helen Weng, Ayuko Hill, Nicole Hay, Christopher Webb, Jamie Cheng, and Erdrin Azemi. 2024. Pre-Trained Foundation Model representations to uncover Breathing patterns in Speech. *arXiv preprint arXiv:2407.13035* (2024).

[42] Vikramjit Mitra, Zifang Huang, Colin Lea, Lauren Tooley, Panayiotis Georgiou, Sachin Kajarekar, and Jefferey Bigham. 2021. Analysis and Tuning of a Voice Assistant System for Dysfluent Speech. In *Interspeech*. https://arxiv.org/pdf/2106.11759.pdf

[43] Payal Mohapatra, Akash Pandey, Bashima Islam, and Qi Zhu. 2022. Speech disfluency detection with contextual representation and data distillation. In *Proceedings of the 1st ACM international workshop on intelligent acoustic systems and applications*. 19–24.

[44] Sofia Monteiro, Ana Fred, and Hugo Plácido da Silva. 2022. Detection of Heart Sound Murmurs and Clinical Outcome with Bidirectional Long Short-Term Memory Networks. In *2022 Computing in Cardiology (CinC)*, Vol. 498. IEEE, 1–4.

[45] Jingping Nie, Yigong Hu, Yuanyuting Wang, Stephen Xia, and Xiaofan Jiang. 2020. SPIDERS: Low-cost wireless glasses for continuous in-situ bio-signal acquisition and emotion recognition. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 27–39.

[46] Jingping Nie, Ran Liu, Behrooz Mahasseni, and Vikramjit Mitra. 2024. Model-Driven Heart Rate Estimation and Heart Murmur Detection Based On Phonocardiogram. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[47] Jorge Oliveira, Francesco Renna, Paulo Costa, Marcelo Nogueira, Ana Cristina Oliveira, Andoni Elola, Carlos Ferreira, Alipio Jorge, Ali Bahrami Rad, Matthew Reyna, et al. 2022. The CirCor DigiScope Phonocardiogram Dataset. *version 1.0. 0* (2022).

[48] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

[49] Parkinson's UK. n.d.. Speech and Communication Problems. https://www.parkinsons.org.uk/information-and-support/speech-and-communication-problems. Accessed: 2024-11-09.

[50] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 914–921.

[51] Nan Bernstein Ratner and Brian MacWhinney. 2018. Fluency Bank: A new resource for fluency research and practice. *Journal of fluency disorders* 56 (2018), 69–80.

[52] Yanzhi Ren, Zhourong Zheng, Hongbo Liu, Yingying Chen, Hongwei Li, and Chen Wang. 2021. Breathing sound-based exercise intensity monitoring via smartphones. In *2021 international conference on computer communications and networks (ICCCN)*. IEEE, 1–10.

[53] BM Rocha, Dimitris Filos, L Mendes, Ioannis Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, P Natsiavas, Ana Oliveira, C Jácome, A Marques, et al. 2018. A respiratory sound database for the development of automated classification. In *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*. Springer, 33–37.

[54] Amrit Romana, John Bandon, Matthew Perez, Stephanie Gutierrez, Richard Richter, Angela Roberts, and Emily Mower Provost. 2021. Automatically Detecting Errors and Disfluencies in Read Speech to Predict Cognitive Impairment in People with Parkinson's Disease.. In *Interspeech*.

[55] Amrit Romana, Kazuhito Koishida, and Emily Mower Provost. 2024. Automatic Disfluency Detection from Untranscribed Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024), 1–14. https://doi.org/10.1109/TASLP.2024.3485465

[56] Amrit Romana, Minxue Niu, Matthew Perez, and Emily Mower Provost. 2024. FluencyBank Timestamped: An updated data set for disfluency detection and automatic intended speech recognition. *Journal of Speech, Language, and Hearing Research* (2024), 1–13.

[57] Johannes Scherr, Bernd Wolfarth, Jeffrey W Christle, Axel Pressler, Stefan Wagenpfeil, and Martin Halle. 2013. Associations between Borg's rating of perceived exertion and physiological measures of exercise intensity. *European journal of applied physiology* 113 (2013), 147–155.

[58] Olabanji Shonibare, Xiaosu Tong, and Venkatesh Ravichandran. 2022. Enhancing asr for stuttered speech with limited data using detect and pass. *arXiv preprint arXiv:2202.05396* (2022).

[59] James M Smoliga, Zahra S Mohseni, Jeffrey D Berwager, and Eric J Hegedus. 2016. Common causes of dyspnoea in athletes: a practical approach for diagnosis and management. *Breathe* 12, 2 (2016), e22–e37.

[60] Arpan Srivastava, Sonakshi Jain, Ryan Miranda, Shruti Patil, Sharnil Pandya, and Ketan Kotecha. 2021. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. *PeerJ Computer Science* 7 (2021), e369.

[61] David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[62] Vernier Software & Technology. n.d.. Go Direct® Respiration Belt. https://www.vernier.com/product/go-direct-respiration-belt/. Accessed: 2024-11-09.

[63] Ronam Toledo, Marcelo R Dias, Ramon Toledo, Renato Erotides, Daniel S Pinto, Victor M Reis, Jefferson S Novaes, Jeferson M Vianna, and Katie M Heinrich. 2021. Comparison of physiological responses and training load between different CrossFit® workouts with equalized volume in men and women. *Life* 11, 6 (2021), 586.

[64] Michigan State University. 2023. Hey Siri, It's Time to Understand Stuttering. https://psychology.msu.edu/news-events/news/archives/2023/hey-siri-its-time-to-understand-stuttering.html. Accessed: 2024-11-09.

[65] Baiyang Wang and Haiyan Zhu. 2022. The recognition method of athlete exercise intensity based on ECG and PCG. *Computational and Mathematical Methods in Medicine* 2022, 1 (2022), 5741787.

[66] Frederik Wiehr, Felix Kosmalla, Florian Daiber, and Antonio Krüger. 2016. Interfaces for assessing the rated perceived exertion (rpe) during high-intensity activities. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 851–855.

[67] Tong Xia, Dimitris Spathis, J Ch, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, et al. 2021. COVID-19 sounds: a large-scale audio dataset for digital respiratory screening. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*.

[68] Chenhan Xu, Tianyu Chen, Huining Li, Alexander Gherardi, Michelle Weng, Zhengxiong Li, and Wenyao Xu. 2022. Hearing heartbeat from voice: Towards next generation voice-user interfaces with cardiac sensing functions. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 149–163.

[69] Baichen Yang, Qingyong Hu, Wentao Xie, Xinchen Wang, Wei Luo, and Qian Zhang. 2023. PDAssess: A Privacy-preserving Free-speech based Parkinson's Disease Daily Assessment System. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 251–264.

[70] Soojeong Yoo, Christopher Ackad, Tristan Heywood, and Judy Kay. 2017. Evaluating the actual and perceived exertion provided by virtual reality games. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 3050–3057.