Poster Abstract: Listen and Then Sense: Vibration-based Sports Crowd Monitoring by Pre-training with Public Audio Datasets

Yen Cheng Chang
yencheng@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Jesse Codling
University of Michigan
Ann Arbor, Michigan, USA

Yiwen Dong
Stanford University
Stanford, California, USA

Jiale Zhang
University of Michigan
Ann Arbor, Michigan, USA

Jeffrey Shulkin
University of Michigan
Ann Arbor, Michigan, USA

Hugo Latapie
Cisco Systems, Inc.
San Jose, California, USA

Carlee Joe-Wong
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Hae Young Noh
Stanford University
Stanford, California, USA

Pei Zhang
University of Michigan
Ann Arbor, Michigan, USA

Figure 1. Crowd reactions at Stanford Maples Pavilion.

Abstract

This paper addresses challenges in monitoring human behavior in crowds through floor vibration sensing, overcoming limitations like subjective manual observation, visual occlusions, and audio interference. Our approach involves tackling limited-data vibration signal tasks by conducting pre-training across modalities, leveraging publicly available audio datasets. By leveraging self-supervised representation learning to pre-train on publicly available audio datasets, our approach reduces data requirements, improves robustness, and minimizes the need for human labeling efforts. Evaluation using in-game stadium vibration data with YouTube audio dataset demonstrates up to 5.8× error reduction for crowd behavior.

Keywords: crowd monitoring, floor vibration, sports game

1 Introduction

Floor vibration sensing emerges as a cost-efficient and privacy-friendly alternative, facilitating continuous and fine-grained monitoring [5]. However, the physical nature of applications involving vibration makes data collection and labeling a significant limitation to the accuracy and robustness of these applications. Although transfer learning or domain adaptation methods have been proposed to tackle these challenges, they require existing and similar datasets, which are often unavailable.

This paper introduces a novel approach that pre-trains the vibration model using a publicly available audio dataset. Our approach first pre-trains a model using a publicly available dataset through a modified Audio-MAE [4] approach, which is inspired by Masked Autoencoders (MAE) [3]. Then, we fine-tune the model using vibration signals from our deployment for classification tasks. Motivated by the challenge of limited labeled training data, we employ a cross-domain training approach, leveraging the abundance of existing out-of-domain unlabeled audio datasets.

For evaluation, we utilize our previous deployment at the Stanford Maples Pavilion for crowd monitoring [1]. This work learns latent representations of game progress and facility layout, integrating this information with vibration data to achieve accurate crowd behavior estimation. Challenges persist in dealing with diverse labeled vibration signals and practical sensor deployment. The proposed adaptation methodology aims to address these challenges, enhancing the efficiency of crowd monitoring.

Our contributions include (1) cross-domain training with different modalities, (2) adaptation to a stadium domain, and (3) evaluation in a real-world stadium setting. Focusing on crowd monitoring as an example dataset, we evaluate the effectiveness of our approach in improving sensing performance with limited training data.

2 Leveraging Public Audio Dataset

While audio and vibration data share physical principles, the sampling and data collection of these two modalities are very different (i.e. vibration sampling rate and bandwidth are lower). To incorporate public audio datasets, we downsample the audio data to align it with the sampling rate of

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00043

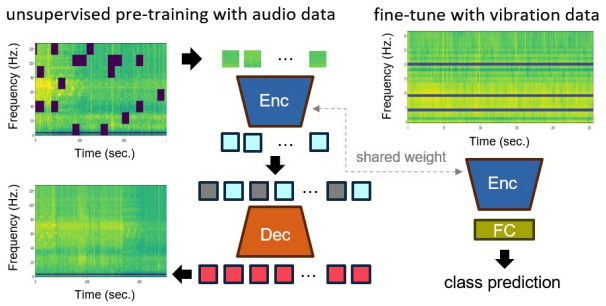

Figure 2. The out-of-domain training procedure is employed for classifying vibration signals.

vibration signals. Our approach follows a two-step process as illustrated in Figure 2: 1) *Unsupervised pre-training with audio data*, and 2) *Fine-tuning with vibration data*.

Unsupervised pre-training with audio data: To accommodate the lower resolution of the vibration signal, we reduce the patch size [4] during the masking procedure. In this step, we pre-train the model by transforming audio signals into spectrograms in an unsupervised manner, using the adopted audio dataset and model parameters.

Fine-tuning with vibration data: To adapt the pre-trained audio model for vibration classification, we fine-tune the model by transforming vibration signals into spectrograms with the same parameters mentioned earlier. Using these spectrograms as input, we fine-tune the model as a multi-class classification model by tuning the fully-connected layer [4] through supervised learning.

3 Real-World Evaluation

Deployment. The experiment was conducted at the Stanford Maples Pavilion, utilizing a total of 12 sensors, with 6 placed inside and 6 outside the stadium. The vibration dataset comprised 3,224 instances, each representing a 1-minute waveform, split into 3,000 for training and 224 for validation. For each waveform, there are 8 different predicted classes of human behavior, including none, three types of clapping based on amplitude, yelling, moving, active, and quiet behaviors.

Sampling rate and Patch Size Choosing. For public audio datasets, we used the AudioSet-2M [2], which contains 2 million 10-second YouTube clips. To align the audio dataset with our vibration signal's lower sampling rate, we downsampled the audio data to 1 kHz. Additionally, due to this lower sampling rate, we adopted a 128×32 spectrum size, which is smaller than that of Audio-MAE. For patch embedding, we employed convolutional kernels with a (4, 4) patch size and non-overlapping strides in time and frequency to avoid shortcuts through overlap in self-supervision.

Comparison with Different Pre-train. The experiment compared vibration pre-training against no pre-training, using 21,840 instances of vibration signals and an equivalent amount of AudioSet-2M data for audio pre-training. Each instance, comprising 10-second recordings, was downsampled to 1 kHz to match the lower sampling rate of vibration signals. We compared our results with no pre-training and pre-training with vibration data only.

The results demonstrate that audio-only pre-training outperforms other modalities by up to 60% or a 5.8X reduction in error. These findings confirm that incorporating cross-modality public audio datasets can significantly improve results in real-world vibration-based applications.

Table 1. Comparison of Accuracy with Different Pre-train.

Pre-trained	None	Vibration	Audio
Accuracy	27.39%	39.97%	**87.51%**

Acknowledgments

This work was funded by Cisco Systems, Inc. The views expressed by the authors do not necessarily represent the official policies of any university or corporation.

References

[1] Yiwen et al. Dong. 2023. GameVibes: Vibration-based Crowd Monitoring for Sports Games through Audience-Game-Facility Association Modeling. In *BuildSys*. 177–188.
[2] Jort F et al. Gemmeke. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*. IEEE, 776–780.
[3] Kaiming et al. He. 2022. Masked autoencoders are scalable vision learners. In *CVPR*. 16000–16009.
[4] Po-Yao et al. Huang. 2022. Masked Autoencoders that Listen. In *NeurIPS*.
[5] Sonu et al. Lamba. 2017. Crowd monitoring and classification: a survey. In *ICCCCS 2016*. 21–31.

Received 1 March 2024; revised 8 March 2024; accepted 15 March 2024