CaNRun: Non-Contact, Acoustic-based Cadence Estimation on Treadmills using Smartphones

Ziyi Xuan*
Columbia University
New York, NY, USA
zx2420@columbia.edu

Ming Liu*
Columbia University
New York, NY, USA
ml4802@columbia.edu

Jingping Nie
Columbia University
New York, NY, USA
jn2551@columbia.edu

Minghui Zhao
Columbia University
New York, NY, USA
mz2866@columbia.edu

Stephen Xia
Columbia University
New York, NY, USA
stephen.xia@columbia.edu

Xiaofan Jiang
Columbia University
New York, NY, USA
jiang@ee.columbia.edu

ABSTRACT

Running with a consistent cadence (number of steps per minute) is important for runners to help reduce risk of injury, improve running form, and enhance overall bio-mechanical efficiency. We introduce CaNRun, a non-contact and acoustic-based system that uses sound captured from a mobile device placed on a treadmill to predict and report running cadence. CaNRun obviates the need for runners to utilize wearable devices or carry a mobile device on their body while running on a treadmill. CaNRun leverages a long short-term memory (LSTM) network to extract steps observed from the microphone to robustly estimate cadence. Through an 8-person study, we demonstrate that CaNRun achieves 96.8 % cadence detection accuracy without calibration for individual users, which is comparable to the accuracy of the Apple Watch despite being non-contact.

KEYWORDS

Acoustic Sensing, Embedded Systems, Cadence Estimation

ACM Reference Format:
Ziyi Xuan, Ming Liu, Jingping Nie, Minghui Zhao, Stephen Xia, and Xiaofan Jiang. 2023. CaNRun: Non-Contact, Acoustic-based Cadence Estimation on Treadmills using Smartphones. In *Cyber-Physical Systems and Internet of Things Week 2023 (CPS-IoT Week Workshops '23), May 09–12, 2023, San Antonio, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3576914.3589561

*These authors contributed equally to this work.

CPS-IoT Week Workshops '23, May 09–12, 2023, San Antonio, TX, USA

ACM ISBN 979-8-4007-0049-1/23/05…$15.00
https://doi.org/10.1145/3576914.3589561

1 INTRODUCTION

A runner's cadence, or step rate, is one of the many metrics that are important for reducing injury, improving speed, and increasing endurance. To estimate and record cadence, a person typically needs to use a wearable (e.g., sports watch) or carry their mobile phone. Close to 90 % of Americans own a smartphone, but only around 30 percent of the population use wearable devices [1]; and carrying a mobile phone can be unwieldy while running. However, we recognize that it is not uncommon to see people place their mobile devices on the treadmill, often to provide entertainment, while running.

We propose CaNRun, a non-contact, mobile, and acoustic-based system for estimating a runner's cadence on a treadmill. CaNRun estimates cadence using sounds observed from the mobile phone's microphone as it sits on the treadmill, leveraging a long short-term memory-based (LSTM) method to extract and count running steps. We leverage an audio-based approach instead of using the phone's camera, as it is less privacy intrusive and requires less computation than vision-based methods. We demonstrate through an 8-person study that CaNRun estimates cadence with more than 96.8 % accuracy, which is comparable to an Apple Watch. CaNRun also provides preliminary ground contact time (GCT) information. Currently, CaNRun runs on a user's smartphone. We also envision that CaNRun can come equipped in future treadmills. We make the following contributions:

- We propose CaNRun, a smartphone-based system that measures a runner's cadence on a treadmill. CaNRun estimates the user's step rate using the audio observed from the smartphone's microphone as it sits on the treadmill.

- We propose an LSTM-based method to clean and extract steps from audio observed from the microphone. This method also provides preliminary insights for the GCT.
- Through an 8-person study, we demonstrate that CaNRun estimates cadence with a 96.8% accuracy without calibration for individual users, which is comparable to the accuracy of an Apple Watch (98.0 %).

2 RELATED WORKS

Cadence is defined as simply the number of steps a person takes in a minute, and is a common metric runners use to improve their consistency. Previous works have found that running cadence has a profound effect on the overall running efficiency (RE), determining up to 28% of an individual's RE, with the closely related stride length metric making up another 23% of RE [8]. In addition to wasting energy, running with a sub-optimal cadence can also cause injuries due to higher impact forces acting on the feet [9].

There is a large market of wearable devices that can estimate running cadence using inertial measurement unit (IMU) data, and these devices are generally fairly accurate when evaluated in control conditions [6]. The downside of wearable devices is that they often require a calibration period and health data from a person to achieve the best results. Additionally, some devices only offer cadence or ground contact time information when connected to GPS, making them ineffective when running on the treadmill. The attachment of wearable devices to the user may also cause discomfort during running, particularly during high-intensity training.

Acoustic systems leveraging both air and ground propagated vibrations and neural networks have also been used to localize occupants [3]. Additionally, previous works have analyzed footsteps and gait using acoustic signals [2].

A closely related metric to cadence is **Ground Contact Time (GCT)**, which is defined as the amount of time a runner's foot spends in contact with the ground. GCT has been found to correspond greatly with running efficiency as well, with a longer GCT being more efficient for distance running, while shorter GCT is more desirable for speed running [5].

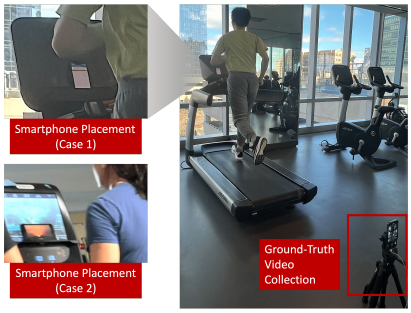


Figure 1: The experiment setup with two smartphone placement cases.

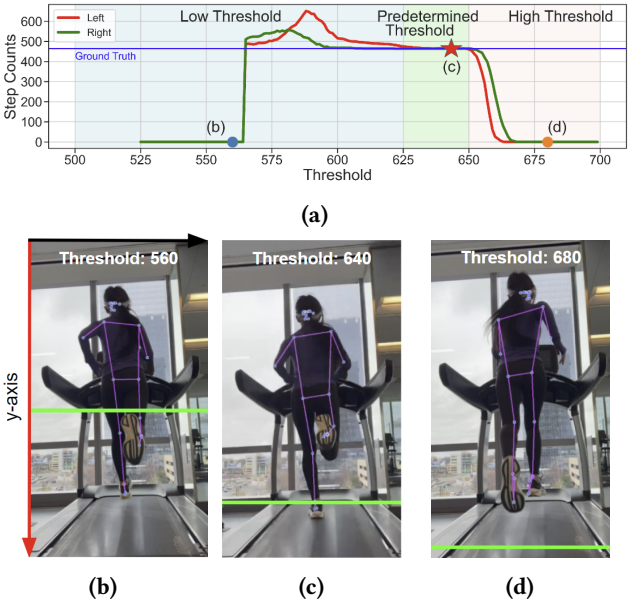


Figure 2: (a): The effects of different threshold settings based on `MediaPipe Pose`. (b)-(d): The visualization of each threshold overlaid on the full-body segmentation mask.

3 STUDY DESIGN

3.1 Experiment Setup

To collect training data and evaluate CaNRun, we recruited 8 voluntary participants, including 4 men and 4 women between ages 18 and 30. The participants' height ranges from 5'1" to 6'1" and their weight falls between 97 lbs and 183 lbs. All procedures of this study was approved by the Columbia University Institutional Review Board (IRB). Every participant wore an Apple Watch for a baseline comparison, placed their smartphone on the treadmill console, and completed four separate running sessions at constant speeds of 5, 6, 7, and 8 miles per hour (*mph*). Each running session lasts for 5 minutes.

As shown in Figure 1, to capture the ground truth of the runners' movements, one video recording device is positioned at the back and center of the treadmill. A smartphone is placed in portrait (case 1) or horizontal orientations (case 2) on the treadmill platform to record running sound data. During each running session, the user activates the Apple Watch's cadence estimator while the video recording devices and smartphone begin capturing video and audio, respectively. After each running session, the audio and video clips are synchronized to obtain ground truth labels.

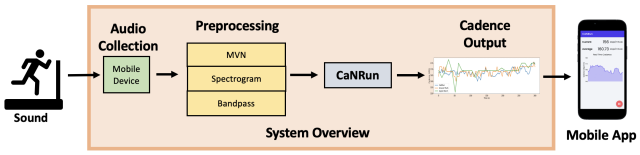


Figure 3: CaNRun's system architecture.

3.2 Video-based Ground-truth Generation

Manually annotating ground-truth footsteps based on video recording at 30 fps is time-consuming. We designed and deployed an approach based on `MediaPipe Pose` to provide automatic ground-truth labeling at decent accuracy, with only a small amount of manual labeling required.

The landmark model in `MediaPipe Pose` can detect 33 body landmarks [7]. However, some important running pose landmarks (e.g., ankle, foot, and heel) tend to overlap from the side view. Hence, ground-truth videos are collected at the back (see Figure 1).

Within each frame of the video, `MediaPipe Pose` generates Y-coordinates for ankle positions. According to our observations, the subject's ankles usually hit a relatively stable horizontal line when in contact with the treadmill belt. Therefore, a predetermined threshold for the ankle coordinates would be ideal for determining the start and end of each step accurately, which identifies both cadence and GCT.

Figure 2 illustrates how different thresholds perform on the video frames of one example runner. When setting the threshold, if positioned too high or too low ankles will either never cross, or the Y-coordinates generated by `MediaPipe Pose` becomes unreliable (see the areas shaded in blue and red in Figure 2a and Figures 2b and 2d).

A 10-second sample video clip from each running session is manually labeled, which is then used to adjust the threshold within the *Predetermined Threshold* region shown in Figure 2a. As such, the threshold that minimizes the error between the manual labels and labels generated by MediaPipe for each running session can be identified (indicated by the red star in Figure 2a and Figure 2c). The results of this proposed ground-truth labeling is evaluated in Section 5.1.

4 SYSTEM ARCHITECTURE

There are several challenges in detecting running footsteps using audio from a smartphone resting on top of the treadmill. Treadmills are usually in noisy environments such as gyms, where surrounding runners generate their own noises and music may be playing. Treadmills are also inherently noisy machines, making it difficult to discern footsteps accurately using naive methods, such as detecting peaks in energy. In addition to ambient noise, the phone may often be placed in different orientations (see Figure 1) and vibrates on the treadmill as the person runs, generating even more noise. The sound produced by the runner is dependent on their physical attributes, such as height and weight. Furthermore, running sounds are also affected by how the user runs (e.g., heel vs. toe impact running, low speed vs. high speed). Figure 5 shows the variances when different subjects run at differing speeds. To account for these sources of error and noise, we propose an LSTM-based method for extracting footsteps from audio collected from the smartphone, which we detail next. CaNRun's full system architecture is shown in Figure 3.

4.1 Preprocessing

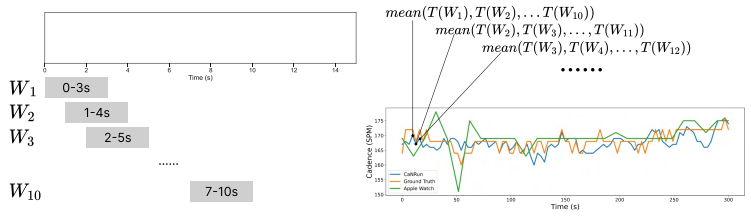
We preprocess audio from the smartphone using mean variance normalization (MVN) before computing the spectrogram and converting to frequency domain, as shown in Figure 5. We see that high and low-frequency noise was common in the signal due to the mobile phone vibrating against the treadmill, and the tread friction noise respectively. However, the mid-frequency bands were relatively clear of high-energy signals apart from footstep impacts. Additionally, we see that the footsteps in the middle-frequency bands and the noise in the low and high-frequency bands have similar levels of energy. As a result, we found that amplitude-based filtering such as spectral gating noise reduction often could not distinguish footsteps from treadmill noises. Hence, we apply a band-pass filter to reduce high and low-frequency noise, while retaining the middle frequencies where footsteps are more apparent.

4.2 Cadence Detection Method

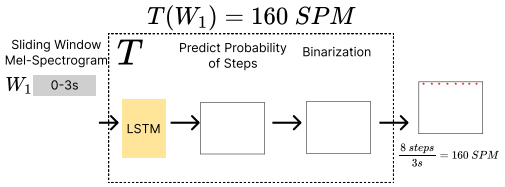
Figure 4 highlights CaNRun's cadence detection algorithm. We use a many-to-many bidirectional LSTM deep learning network to extract footsteps from audio. We leverage an LSTM because audio is a form of time series data. The input to the model is an n-second window of the spectrogram, and the output at each time point is the probability of the user's foot being in contact with the treadmill (e.g., a footstep is occurring), which provides information for both cadence and GCT. Next, we threshold (binarize) the probabilities and use a running average to smooth the output time series of probabilities into binary labels. To obtain the number of steps within the n-second window, we search for and count peaks within the time series of binarized labels. The cadence, or steps per minute (SPM) then results from normalizing the counted number of steps by the window size (n seconds) and extrapolating to the minute-scale, as shown in Equation 1.

$$Cadence(SPM) = counted\_steps \times \frac{60}{n} \quad (1)$$

The next concern is determining the window size n. Smaller window sizes have the advantage of immediate feedback for the runner, at the cost of actual cadence resolution and more variation in the output. For example, if the chosen window



(a) Cadence estimation pipeline. CaNRun 1) takes the preprocessed data and slice it into three-second windows; 2) estimates cadence for each window using the operation $\boldsymbol{T}$ defined in (b), and 3) calculates the moving average for 10 windows for one prediction.



(b) Cadence estimation algorithm $\boldsymbol{T}$.

Figure 4: CaNRun's detection algorithm.

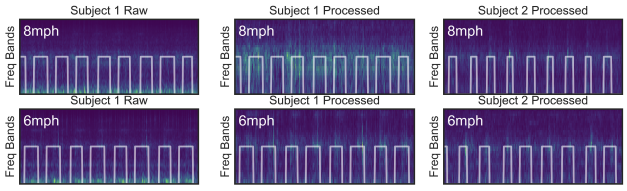


Figure 5: The illustration of the effects of the proposed preprocessing mechanism ((a) and (c)) and comparison of two subjects' spectrogram at different running speeds ((b) and (c)). The ground-truth footsteps are superposed for visualization purposes.

starts or ends during the middle of a footstep, the estimated cadence may contain one extra or one less footstep, which would significantly affect the estimated cadence if the window size was small, but have less impact if the window size is large. However, a larger window size may require a larger LSTM model to process, which increases latency and computation. To obtain the best of both worlds, we decide to use a small window size of $n = 3$ seconds with a 1-second window shift. However, to minimize the impact of partial footsteps in any window, we average the cadence estimated from 10, $n = 3$ second, windows. In other words, CaNRun's estimation of cadence occurs over a longer 10-second period. In this way, as shown in Figure 4, we can reuse the smaller LSTM-based extractor for each $3-second$ window to reduce computation, while being robust to artifacts arising from footsteps being partially cutoff at the start or end of each window. We decide to average over a 10-second period because this is the same period of time that the Apple Watch uses to estimate and report cadence.

4.3 Smartphone Platform

Figure 6 shows the user interface for CaNRun's smartphone system. CaNRun samples audio from the microphone at 22.05kHz. We implemented CaNRun on an iPhone 14 Pro and implemented our LSTM-based footstep extractor (Section 4.2) using TensorFlow Lite [4]. For one 3-second clip of audio, it takes 10.8 *ms* to preprocessing the audio signal, and

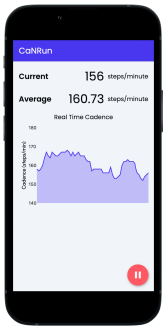


Figure 6: Smartphone platform with historical and current cadence displayed to the end user in real-time.

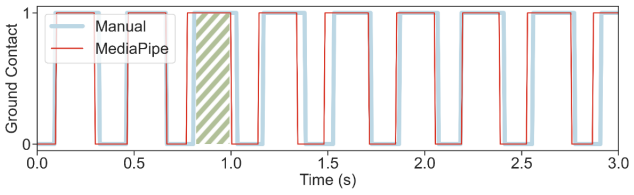


Figure 7: `MediaPipe` labels overlaid manual labels with accuracy of 99.35 %.

45.1*ms* for model inference and cadence estimation. Since we use a 1-second window shift, we update the real-time cadence once per second using a rolling average across a 10-second period of windows. In total, CaNRun takes 55.9 *ms* to update (e.g., convert to frequency, run LSTM-based extractor, averaging, etc.) its estimate of the user's cadence every three seconds, making it capable of running in real-time.

5 EVALUATION

5.1 Video-based Ground Truth Collection

We evaluate our video-based footstep ground-truth detection pipeline against manual labeling to determine if our MediaPipe-based solution can be used to accurately and automatically generate ground truth labels in place of labor-intensive manual labeling. Figure 7 compares the periods of

Speed (mph)	Subject 1 Error	Subject 2 Error	Subject 3 Error	Subject 4 Error	Subject 5 Error	Subject 6 Error	Subject 7 Error	Subject 8 Error
5	99.45 %	99.67 %	99.46 %	97.86 %	98.93 %	99.48 %	99.48 %	99.28 %
6	99.74 %	97.10 %	98.54 %	99.74 %	98.47 %	99.22 %	99.12 %	99.83 %
7	99.33 %	97.77 %	99.72 %	97.77 %	98.31 %	99.86 %	99.08 %	98.77 %
8	99.20 %	99.45 %	99.86 %	99.80 %	99.48 %	99.35 %	98.88 %	99.18 %

Table 1: Average `MediaPipe` labeling error compared to manual labeling.

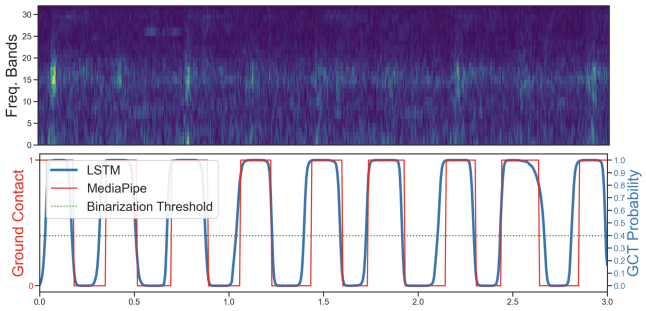


Figure 8: LSTM output overlaid with `MediaPipe` labels and corresponding spectrogram. The LSTM output correspond with the spectrogram peaks, as well as the `MediaPipe` ground-truth labels.

time where our MediaPipe-based solution detects that the foot is contacting the ground (ground contact) compared with manual labeling. Table 1 provides a breakdown for the error across six subjects at all speeds. Overall, our MediaPipe-based solution differs from manual labels by, on average, only 0.65 %. As such, we adopt our video-based solution for labeling training and testing data to train and evaluate CaNRun's cadence estimation method.

5.2 System Performance Evaluation

To evaluate CaNRun, we use the data collected from subjects 1 to 4 (2 male and 2 female) and gathered from subjects 5 to 8 (2 male and 2 female) as the training and testing data, respectively. As mentioned in Section 3.1, the cadences are calculated for 5-minute running sessions at 4 different speeds. We compare CaNRun against the Apple Watch.

Figure 8 shows the CaNRun's LSTM output from 3-second audio from one subject, which generates the probability that the user's foot is in contact with the treadmill (i.e., GCT) and highly correlates to the ground-truth generated by the MediaPipe-based method. The green dashed line indicates the binarization threshold for cadence estimation (empirically set at 0.4). Figure 9 shows that the time series of the estimated cadence for one example running session between CaNRun, the Apple Watch, and the ground truth match very closely. This shows CaNRun closely follows the ground truth and has a comparable performance as Apple Watch.

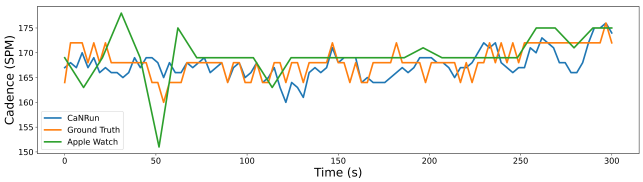


Figure 9: Comparison of CaNRun with `MediaPipe`-based ground truth and Apple Watch on cadence for one 5-minute running session when a subject runs at 6 mph.

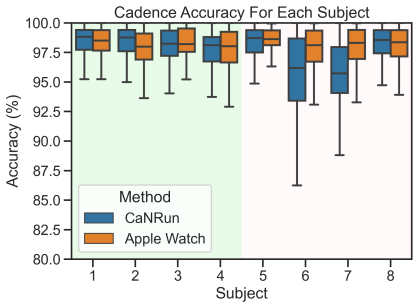


Figure 10: Cadence estimation accuracy distribution for each subject (green shaded portion: from training data, red shaded portion: from testing data). We can see that CaNRun's performance is fairly comparable to the Apple Watch, even for unseen participants.

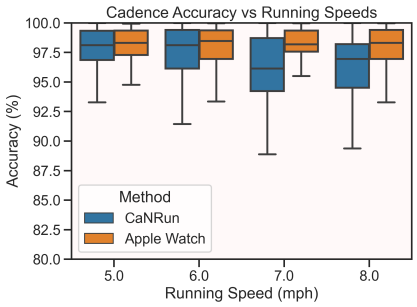


Figure 11: Cadence error for different running speeds.

	Training Data	Testing Data
CaNRun	**97.79 %**	**96.77 %**
Apple Watch	97.93 %	98.04 %

Table 2: CaNRun's average cadence estimation accuracy on training and testing data.

Figure 10 and Figure 11 show the breakdown in cadence estimation accuracy over each subject and running speed, respectively. CaNRun has comparable performance to the Apple Watch, but does not require the user to wear anything. As shown in Table 2, CaNRun achieves 96.77 % cadence estimation accuracy on unseen participants, which is comparable to the estimation performance of Apple Watch. *Note*

that Apple Watch's cadence detection results are calibrated for individual users based on their outdoor running history.

In addition, as shown in Figure 8, the output of the LSTM-based detector is a probability of the GCT before applying the empirical binarization threshold for cadence estimation. Though not the primary focus of this paper, the GCT calculated from the LSTM-based detector on the testing data is 90.5 % accurate, which is high enough to enable future algorithm design for accurate GCT estimation.

6 DISCUSSION

CaNRun is a first step towards robust non-intrusive monitoring of running form. We plan to explore the following directions in future work.

1. Larger user study and more equipment variety. Although our cadence estimation error was comparable to the Apple Watch, we only studied a small number of participants on a single type of treadmill. We plan to conduct larger user studies across multiple treadmills and environments in future work, as well as possibly implement more advanced audio filtering systems like AVA [11] [12] to isolate footsteps from noise more effectively while preserving privacy [10].

2. Detecting other metrics for improving running form. In this work, we focus on estimating cadence with preliminary findings for ground contact time (GCT). We plan to further explore the algorithms for robust GCT detection.

3. Classifying running type. We observed that there are two distinct types of running on the treadmill, which we call "push" and "pull", which can significantly impact performance. The former is when the runner pushes off the treadmill at the end of the step, causing the sound to lag slightly behind the initial ground contact. Pull running is when the runner pulls himself forward at the initial impact, resulting in the sound occurring at the beginning of ground contact. We plan on classifying these styles in future work.

7 CONCLUSION

We propose CaNRun, a mobile acoustic-based running cadence estimator on a treadmill by using the footstep extractor based on LSTM. CaNRun is deployed and tested on a smartphone system to realize real-time cadence detection. CaNRun achieves an accuracy of 96.8 % without requiring individual calibration. We also show that CaNRun has a potential ground contact time estimation capability with 90.5 % accuracy based on preliminary results.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation under Grant Numbers CNS-1704899, CNS-1815274, CNS-1943396, CNS-1837022, and CMMI-2218809. 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] 2022. Wearable Healthcare Technology Statistics. https:https://vicert.com/blog/wearable-healthcare-technology-statistics/.

[2] M. Umair Bin Altaf, Taras Butko, Biing-Hwang Fred Juang, and Biing-Hwang Fred Juang. 2015. Acoustic Gaits: Gait Analysis With Footstep Sounds. *IEEE Transactions on Biomedical Engineering* (2015). https://doi.org/10.1109/tbme.2015.2410142

[3] Chao Cai, Henglin Pu, Peng Wang, Zhe Chen, and Jun Luo. 2021. We hear your pace: Passive acoustic localization of multiple walking persons. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–24.

[4] Robert David, Jared Duke, Advait Jain, Vijay Janapa Reddi, Nat Jeffries, Jian Li, Nick Kreeger, Ian Nappier, Meghna Natraj, Tiezhen Wang, et al. 2021. Tensorflow lite micro: Embedded machine learning for tinyml systems. *Proceedings of Machine Learning and Systems* 3 (2021), 800–811.

[5] Rocco Di Michele and Franco Merni. 2014. The concurrent effects of strike pattern and ground-contact time on running economy. *Journal of science and medicine in sport* 17, 4 (2014), 414–418.

[6] Heontae Kim, Wei Sun, Mary Malaska, Bridget Miller, and Ho Han. 2019. Validation of Wearable Activity Monitors for Real-Time Cadence. In *International Journal of Exercise Science: Conference Proceedings*, Vol. 2. 35.

[7] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[8] Marcus Peikriszwili Tartaruga, Jeanick Brisswalter, Leonardo Alexandre Peyré-Tartaruga, Aluísio Otávio Vargas Ávila, Cristine Lima Alberton, Marcelo Coertjens, Eduardo Lusa Cadore, Carlos Leandro Tiggemann, Eduardo Marczwski Silva, and Luiz Fernando Martins Kruel. 2012. The relationship between running economy and biomechanical variables in distance runners. *Research Quarterly for Exercise and Sport* 83, 3 (2012), 367–375.

[9] Jedd Wellenkotter, Thomas Kernozek, Stacey Meardon, and Timothy Suchomel. 2014. The Effects of Running Cadence Manipulation on Plantar Loading in Healthy Runners. *International journal of sports medicine* 35 (03 2014). https://doi.org/10.1055/s-0033-1363236

[10] Stephen Xia and Xiaofan Jiang. 2020. PAMS: Improving Privacy in Audio-Based Mobile Systems. In *Proceedings of the 2nd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things* (Virtual Event, Japan) *(AIChallengeIoT '20)*. Association for Computing Machinery, New York, NY, USA, 41–47. https://doi.org/10.1145/3417313.3429383

[11] Stephen Xia and Xiaofan Jiang. 2022. AvA: An Adaptive Audio Filtering Architecture for Enhancing Mobile, Embedded, and Cyber-Physical Systems. (2022), 118–131. https://doi.org/10.1109/IPSN54338.2022.00017

[12] Stephen Xia, Jingping Nie, and Xiaofan Jiang. 2021. CSafe: An Intelligent Audio Wearable Platform for Improving Construction Worker Safety in Urban Environments. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (Co-Located with CPS-IoT Week 2021)* (Nashville, TN, USA) *(IPSN '21)*. Association for Computing Machinery, New York, NY, USA, 207–221. https://doi.org/10.1145/3412382.3458267