Low-Cost Gunshot Detection System with Localization for Community Based Violence Interruption

Isaac Manring, James H. Hill
Department of Computer and Information Science
Indiana University - Purdue University Indianapolis, IN
(imanring, hilljh)@iupui.edu

George Mohler
Department of Computer Science
Boston College, Chestnut Hill, MA
mohlerg@bc.edu

P. Jeffrey Brantingham
Department of Anthropology
University of California, Los Angeles, CA
branting@ucla.edu

Thomas Williams
thomasforsythewilliams@gmail.com

Bruce White
b@astrosensor.com

***Abstract*—There is growing interest in U.S. cities to shift resources towards community-led solutions to crime and disorder. However, there is a simultaneous need to provide community organizations with access to real-time data to facilitate decision making, to which only the police normally have access. In this work we present a low-cost gunshot detection system with localization that has been developed for community-based violence interruption. The distributed real-time gunshot detection sensor network is linked to a mobile phone-based alert and tasking system for exclusive use by civilian gang interventionists. Here we present details on the system architecture and gunshot detection model, which consists of an Audio Spectrogram Transformer (AST) neural network. We then combine gradient maps of the input to the AST for time of arrival identification with a Bayesian maximum a posteriori estimation procedure to identify the location of gunshots. We conduct several experiments using simulated data, open data from the commercial ShotSpotter detection system in Pittsburgh, and data collected using our devices during live-fire experiments at the Indianapolis Metropolitan Police Department (IMPD) gun firing range. We then discuss potential applications of the system and directions for future research.**

***Index Terms*—gunshot detection, localization, transformer neural network, violence interruption**

I. Introduction

The murder of George Floyd at the hands of Minneapolis Police officer Derek Chauvin on May 25, 2020, accelerated the call for police reform including the shift of resources away from law enforcement into community-led solutions to crime and disorder [1]. Community-led solutions are thought to be better calibrated to needs because they are implemented by community 'insiders' [2]. They are also thought to be more likely to lead to just outcomes because they often focus on root causes [3], and are not predicated on compliance with the law through threat of force [4].

Community-led solutions to crime and disorder are hampered, however, by poor access to data relevant to designing, implementing and evaluating interventions. Publicly available data about crime and disorder, for example, is generally incomplete (e.g., missing detailed information about offenders and victims), spatially imprecise and stale [5]. Data sourced by word-of-mouth within a community (including traditional and social media) also suffers certain limitations. It may freely mix opinion with factual information [6], and reflects the specific interests of those collecting the data [7]. In other words, community-led solutions are prone to their own set of biases linked to data quality and availability. A shift in resources to community-led solutions may in fact exacerbate problems of crime and disorder unless there is a parallel move to provide community actors access to objective, real-time data to support evidence-based solutions.

In this work we present a low-cost gunshot detection system with localization that has been developed for community based violence interruption. The distributed real-time gunshot detection sensor network is linked to a mobile phone-based alert and tasking system for exclusive use by civilian gang interventionists (see Figure 1). Our ultimate goal is to connect civilian-led violence interruption teams to essential data and provide smart decision-support tools for effective prevention of gun violence.

Our work here improves upon a previous system in two ways [8]. Specifically, the present system achieves increased detection accuracy using a large transformer neural network and incorporate a Bayesian approach for spatial localization. The remainder of this paper is organized as follows. In Section II we review recent literature on gunshot detection and localization on IoT devices. In Section III we present details on the system architecture and gunshot detection transformer model. In Section IV we describe our approach to localization and in Section V we conduct several experiments using simulated data, open data from the commercial ShotSpotter detection system in Pittsburgh, and data collected using our devices

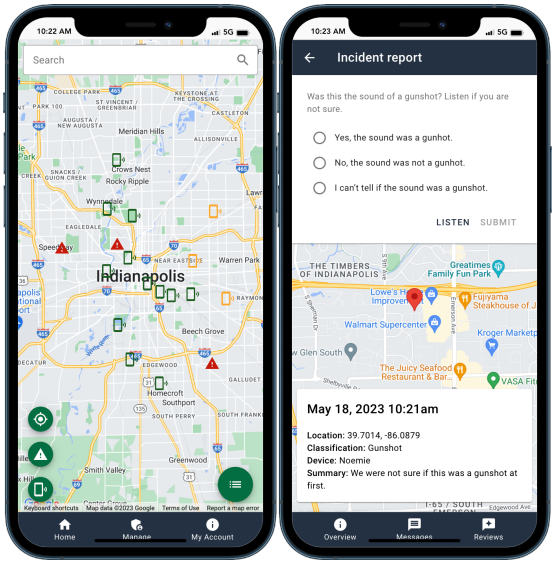


Fig. 1: Community violence interruption mobile application. Left: detected gunshots are shown on a map along with locations of violence interrupters. Right: Violence interrupters also can review gunshot detections for false positives in the application.

during live-fire experiments at the Indianapolis Metropolitan Police Department (IMPD) gun firing range. In Section VI we discuss potential applications of the system and directions for future research.

II. Related Work

In [9], the authors use acoustic gunshot detection with a CNN classifier to monitor forests for poaching. Recently various deep learning architectures have been applied to gunshot detection. In [10], a gunshot classification model is developed using a convolutional-GRU network. In [11], three architectures for gunshot detection on smartphones are compared: two CNN models and one transformer model. The authors found that the transformer-based model was more robust to noise and more accurate than CNN-based models. In [12], a full detection and localization system on a STM32 micro-controller is presented. The authors performed detection with a SVM on MFCC preprocessed audio. They trained on 300 samples collected in live-fire experiments, using a variety of guns. Localization was done with the iterative Levenberg–Marquardt algorithm.

III. System Architecture and Detection Model

The gunshot detection devices are Raspberry Pi 4s in water-proofed cases with SPH0645LM4H-B microphones (see Fig. 2). To detect gunfire the devices continuously run a classification model on the audio input stream. When a classification model predicts a gunshot, a message is sent over the cellular network to the server with the device location, time of detection, and audio clip. If there is a sufficient number of detections, the server will perform localization. The server

Fig. 2: Two low-cost (<$200 U.S.) gunshot detection devices each utilizing a Raspberry Pi 4 and a SPH0645LM4H-B microphone.

then pushes the notification and audio clip to an app, notifying community intervention workers.

Gunshot detection was performed by taking consecutive two second audio clips and classifying them as containing a gunshot or not. We used audio classification model Audio Spectrogram Transformer (AST) [13]. AST applies the mel-spectrogram transformation to the audio before classification to get an image where each column contains the frequencies for a particular time step. The resulting image is fed into a vision transformer (Data Efficient Image Transformer [14]), where it is split into overlapping patches, projected by a convolution, and fed into the transformer encoder blocks (Fig. 3). Finally, the output of the transformer is classified by a dense neural layer. The output has a size of 2 (gunshot, no gunshot). We fine-tuned the model using the initial weights from training on AudioSet after training on ImageNet. We used the hyperparameters from [13] that were optimized on AudioSet.

Our dataset was a conglomeration of publicly available audio data that contained gunshots. We trained on all UrbanSounds8K [15], a subset of Google's AudioSet [16], and on all of Gunshot Audio Forensics Dataset[1]. There were 4,397 audio clips with gunshots in them and 25,747 audio clips without gunshots in them. Even with such a diverse dataset, AST performed well on test data. With a 0.5 threshold, AST achieved an accuracy of 0.986, F1 score of 0.959, recall of 0.969, precision of 0.950, and an AUC of 0.997. The model was quantized to deploy on a Raspberry Pi 4.

IV. Localization

Localization in Wireless Acoustic Sensor Networks (WASN) is generally divided into five approaches based on the features used: Direction Of Arrival, input energy, Time Difference of Arrival, Time of Arrival, and steered response power [17]. Direction Of Arrival (DOA) localization is not

[1]http://cadreforensics.com/audio/

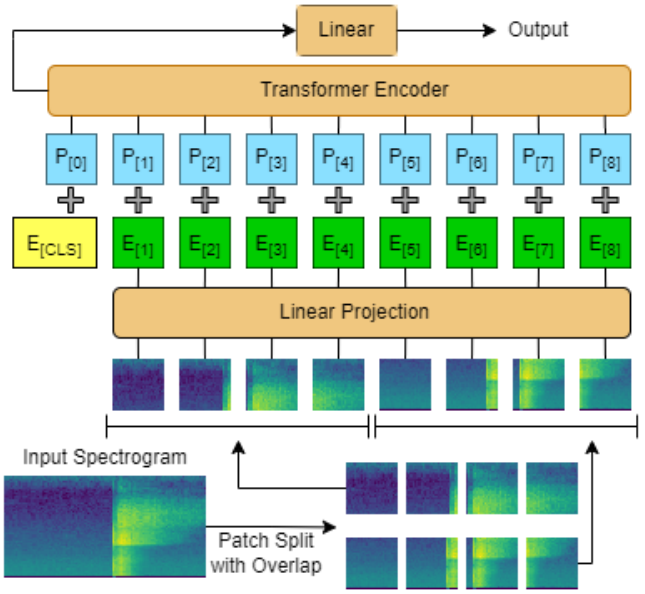


Fig. 3: Audio Spectrogram Transformer Architecture as in [13]

possible for our setup since DOA estimates rely on multiple microphones and our devices only have one microphone per device. Energy-based localization is generally less accurate and requires us to find the energy of the gunshot as opposed to the significant background noise in an urban environment. Time Difference Of Arrival (TDOA) localization using generalized cross-correlation is prone to error in noisy environments like the one in which these devices will be deployed. More importantly, generalized cross-correlation is not particular to which sound it looks at and thus may get the time offset for the wrong sound or use a compromise between two sounds. This is particularly important for a WASN in an urban environment because the acoustic environments of the devices are not necessarily the same. Steered-Power Response methods also use TDOA information. While there are multiple TDOA algorithms, we chose to use Time Of Arrival (TOA) localization. TOA localization simply uses the time of the arrival of the sound at the detectors, unlike the rest of the approaches. Like TDOA localization, TOA localization requires synchronization and accurate TOA estimates.

We formalize the TOA localization problem as follows. Let t_i be the time of arrival of the sound at device i, $\mathbf{x_i}$ be the location of device i, $\mathbf{x_0}$ be the location of the sound source, t_0 be the time of sound emission, and s be the speed of sound. The time of arrival (t_i) is stochastic due to acoustic noise and error in finding the exact TOA in an audio clip. Therefore, TOA localization can be modeled by,

$$t_i|\mathbf{x_0}, t_0, \mathbf{x_i} \sim N(\frac{\|\mathbf{x_i} - \mathbf{x_0}\|}{s} + t_0, \sigma^2) \tag{1}$$

where N is the normal distribution, and σ is the standard deviation. Our goal is to jointly estimate $\mathbf{x_0}$ and t_0. For notational simplicity, we denote the matrix of device locations as X and the vector of TOAs as $\mathbf{t}$. In the Bayesian paradigm,

$$p(\mathbf{x_0}, t_0|X, \mathbf{t}) \propto p(X, \mathbf{t}|\mathbf{x_0}, t_0)p(\mathbf{x_0}, t_0) \tag{2}$$

In TOA localization, the device locations are fixed so that,

$$p(X, \mathbf{t}|x_0, t_0) = p(\mathbf{t}|x_0, t_0, X) \tag{3}$$

We assume a uniform prior on $\mathbf{x_0}$ and t_0 and use the maximum a posteriori estimate for the location. The problem becomes maximizing the log posterior ℓ over $\mathbf{x_0}$ and t_0, where

$$\ell(\mathbf{x_0}, t_0) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\frac{\|\mathbf{x_0} - \mathbf{x_i}\|}{s} + t_0 - t_i)^2 \tag{4}$$

This setup has been used for localization when t_0 was known [18]. We use a Quasi-Newton iterative method for optimization. Since this optimization technique can be sensitive to the starting position, we choose the initial state carefully; the initial $\mathbf{x_0}$ was set to the mean of the device locations detecting the shot. The initial t_0 was set to the minimum TOA minus the time it takes sound to travel from the initial $\mathbf{x_0}$ to the device with the minimum TOA.

A. *Time of Arrival Identification*

The Time Of Arrival for a gunshot can be found by adding the start time of the clip that contains the shot and an offset for when the gunshot was first observed in the audio clip. The difficulty is finding this offset. The CUMSUM change point detection algorithm has been used to estimate TOAs in [19]. This algorithm, however, is unsuitable in our context because of the potential presence of other sound sources that may be as loud or louder than the gunshot itself, causing multiple change points in the audio. An algorithm employing the discriminative power of a classifier is necessary to determine which sound onset to to measure. We investigated two such options.

In [20], TOAs are identified via matched filtering. Here we compare the audio in which we are attempting to identify the TOA to a template audio where the TOA has already been found. We find the time offset that maximizes the cross-correlation of the template and the input. This approach suffers from the same noise problem as the Generalized Cross-correlation for TDOA estimation does. While matched filtering uses a clean template unlike Generalized Cross-correlation, the gunshot template is not necessarily a perfect match for a gunshot detected in the urban environment.

A second option is to use the location in the audio input that the Audio Spectrogram Transformer deemed most important. Due to the architecture of AST, the attention maps from the transformer as in [21] were too course-grained in the time dimension (on the order of 0.1 seconds) to be useful in localization. The gradient map of the input has also been used to determine what a model deems important [22]. We computed the gradient map from [22], summed over the frequencies, and smoothed the result with a convolution kernel. This method provides more fine-grained time information, but has a significant amount of variation associated with it. We noticed that the onset of a gunshot is accompanied by a sharp

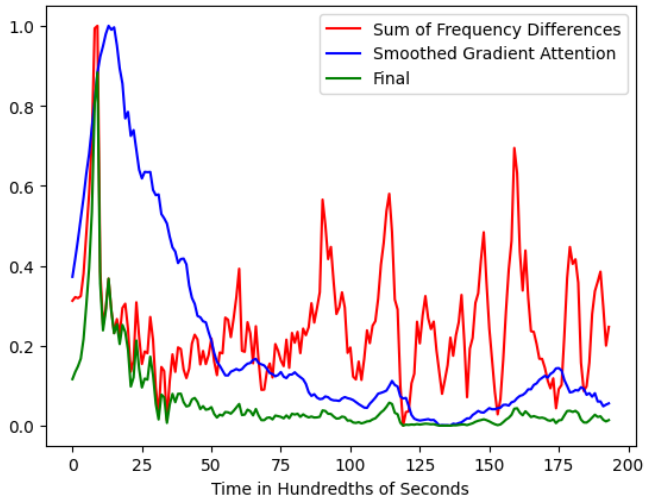


Fig. 4: Normalized signals for Gradient-based TOA Identification. All values have been rescaled to be between 1 and 0. Shot occurred at 0.1 sec

rise in the values of all frequency bins in the spectrogram due to the muzzle blast. We summed the frequency bins at every timestep of the spectrogram and smoothed the result by a convolution. Then we took the difference of this for every timestep and its predecessor to find the exact moment of the rise in air turbulance. We normalized these two measures to be between 1 and 0 and multiplied them together. We took the maximum of this value for an estimate of the TOA. The gradients provide the general location of the TOA in the clip, while the difference in the sum of frequency bins gives the desired precision.

An example is helpful to illustrate this heuristic approach. In Fig. 4, the sound of a gunshot arrives at approximately 0.1 seconds. Someone begins speaking at approximately 0.6 seconds and continues until the end of the clip. This can be observed by the large variation in the turbulence (red). The model correctly identifies the beginning of the clip as containing a gunshot (the blue signal is very high initially). The final signal (green) is only high when both the model deems the section of the audio to be important (blue) and there is an onset of a new sound (red).

Anomaly detection can be used for multiple shot TOA identification for both the matched filtering and model gradient methods.

B. Error Estimation

Error is introduced into the localization system by wave propagation dynamics (through obstructions or multipath propagation) and inaccuracies in TOA estimation. It is therefore useful to characterize the localization estimate with an error region. The error region will not only display the system's confidence in the localization, but it will also help determine a zone of reasonable belief in which to expect the sound source.

In [23], the authors used a Bayesian setup for TDOA localization of acoustic emission in structural integrity monitoring. Using Bayes rule, they found the un-normalized posterior of source location and wave velocity. They then used Markov Chain Monte-Carlo (MCMC) sampling to sample from the posterior and find a normal approximation of the posterior. The approach in [23] differs from the current work in several key aspects: they used TDOA localization rather than TOA localization, they used MCMC sampling instead of Laplace approximation, and they did not attempt to predict an error region using experimental data.

Rather than using MCMC sampling, we chose Laplace approximation becasue it is much more computationally efficient. Laplace approximation is a second order Taylor expansion for approximating an un-normalized posterior with a normal distribution. It uses the maximum a posteriori estimate, which we found in the localization section, and the Hessian of the negative log posterior evaluated at the estimate. The Hessian matrix can be analytically found (see the appendix) and constitutes the precision matrix. Letting $\theta = (\mathbf{x_0}, t_0)$, we then have that

$$\Lambda = \frac{\partial^2}{\partial \theta^2} - \ell(\theta). \tag{5}$$

If we invert the precision, we get the covariance matrix. We are interested in the covariance of the $\mathbf{x_0}$ marginal, which is obtained by ignoring the t_0 dimension of the full covariance. The eigen-decomposition of the covariance matrix provides the variance as the eigenvalue along the associated eigenvector. Let $\mathbf{v}$ be an unit eigenvector of $\Sigma = \Lambda^{-1}$ with associated eigenvalue λ and $\mathbf{x_{MAP}}$ be the maximum a posteriori estimate. Then we have,

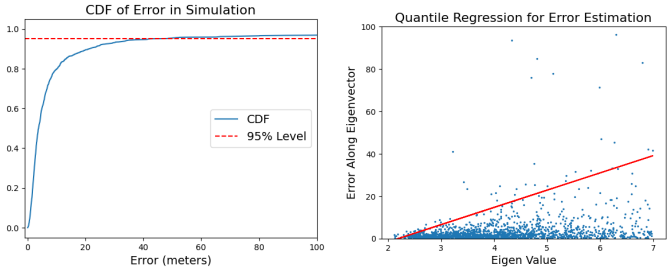
$$Var((\mathbf{x_0} - \mathbf{x_{MAP}}) \cdot \mathbf{v}) = \lambda. \tag{6}$$

Several sources of error can be characterized by σ^2, such as error in device synchronization and TOA identification error. Additionally, error resulting from obstructions and mutlipath propagation directly effect the values of t_i, but only have the potential to increase t_i from the theoretical t_i. Other error sources cannot be captured by σ^2, such as error in the speed of sound constant from temperature or wind. Due to the partial explanation of error by the model, we chose to use the theoretical eigenvalues as predictors for the actual error. The model can be learned from real data collected using the localization system. The regression model not only (indirectly) determines σ^2, but also adds a term for unaccounted sources of error.

Because our ultimate goal is to get a confidence region for a localization, we use quantile regression for the 95% quantile. We predict the 95% quantile of the error projected along an eigenvector with the square root of the eigenvalue because theoretically the 95% quantile should be $1.96\sqrt{\lambda}$. Therefore, the model is,

$$q_{0.95}(\mathbf{error}_i \cdot \mathbf{v_{i,j}}) = \alpha\sqrt{\lambda_{i,j}} + \beta + \epsilon_i, \tag{7}$$

where the subscript i identifies the shot, j identifies the eigenvalue-eigenvector pair, $q_{0.95}(\cdot)$ is the 95% quantile func-



(a) Cumulative Distribution Function of Error from Simulation with 95% quantile

(b) Scatter of eigenvalue versus error with 95% quantile regression line.

Fig. 5: Simulation Results

tion, and ϵ_i is random error. The resulting 95% confidence region is elliptical.

It should be noted that we dropped σ^2 in the calculation of Λ because σ^2 is absorbed by α the in the regression equation. From our datasets, experimental and simulated, we notice that the distribution of eigenvalues is strongly skewed to the right. Further, if two eigenvalues are outliers, the eigenvalues are not good predictors of which error is greater. For this reason, we capped the eigenvalues and made the error region estimation for large eigenvalues based on the sample quantile of error above the chosen cap.

V. Experiments and Results

A. Simulation

In our simulation, we varied the number of devices detecting a shot from 3 to 15. For each number of detecting devices, we simulated 50 shots whose locations were uniformly distributed on a 1000x1000 meter square. For each of those 50 shots we generated new device locations also uniformly distributed on the 1000x1000 square. The TOAs were calculated and then a normally distributed error with $\sigma = 0.01$ seconds was added. Finally, we ran the localization procedure developed above and found the Laplace approximation of the posterior.

In this simulation the median error was 3.8 meters and the error distribution was strongly skewed to the right (see Fig. 5). In practical terms, we can localize a gunshot within 41.8 meters of the true location with 95% confidence. We also fit the quantile regression from equation (7) with a cap of 7 and performed statistical inference on the parameters using the bootstrap method (Fig. 5). The eigenvalue based predictor, $\sqrt{\lambda}$, was found to be a statistically significant estimator at the 0.1% significance level. This indicates that our elliptical error region is more informative than a constant circular region.

B. ShotSpotter Data Set

In 2018 ShotSpotter and the Pittsburgh Bureau of Police conducted live fire experiments in Pittsburgh PA to test gunshot localization [24]. They chose nine different locations called "firing positions" from which gunshots were produced. Using the TOA data and locations provided through this experiment, we performed localization with our proposed method.

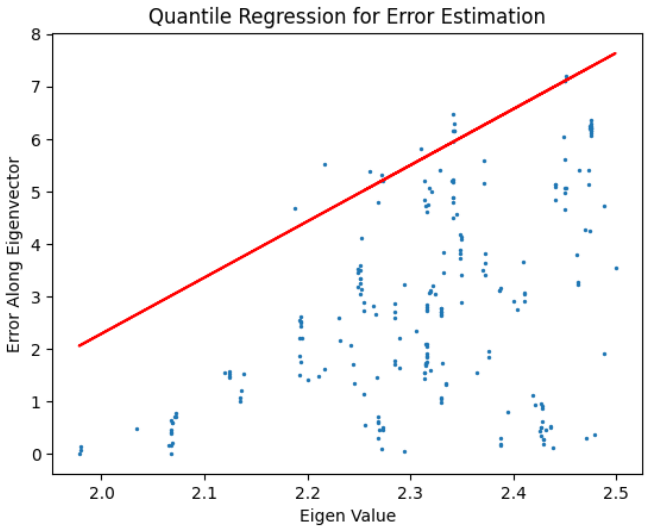


Fig. 6: 95% quantile regression line with eigenvalue-error pairs from SpotSpotter data. The eigenvalue-error pairs show an approximately linear trend in the 95% quantile as expected.

Our approach performs similarly to the methods discussed in [24] (compare TABLE I in this work to TABLE III in [24]).

TABLE I: Joint Optimization Algorithm

Firing Position	ϵ (m)	σ_1	σ_2
FP1	3.43	0.57	0.27
FP2	4.19	1.15	0.21
FP3	2.46	1.29	0.36
FP4	5.70	1.15	0.52
FP5	3.97	1.25	0.36
FP6	5.12	0.84	0.25
FP7	5.93	0.76	0.37
FP8	2.12	1.21	0.36
FP9	5.50	0.57	0.27

Results of our algorithm applied to ShotSpotter experiment data where ϵ is error in meters, measured as the distance from the center of localization estimates to the true shot location for all shots from that location. Here σ_1 and σ_2 are the square roots of the eigenvalues of the covariance matrix for the estimated shot locations.

Additionally, the Laplace approximation for error estimation was useful. After we capped the data, there were 233 data points for the model in equation (7). The eigenvalue estimator $\sqrt{\lambda}$ was a statistically significant estimator at the 0.1% significance level using the bootstrap method. Thus our novel localization error estimate was informative for this data also.

C. IMPD Gun Range Experiments

To further test our methodology, we brought our devices to the IMPD firearms training range where we observed the shots from officers in training. Even though the AST had no prior training on data collected through our device, it still correctly identified 99/116 audio clips as containing a gunshot, significantly better than the 35/280 accuracy reported prior to fine-tuning in [8]. The confusion matrix is shown in Table II.

TABLE II: Confusion Matrix for IMPD Range Data

	Predict	Not Predict
Gunshot	99	17
No Gunshot	16	57

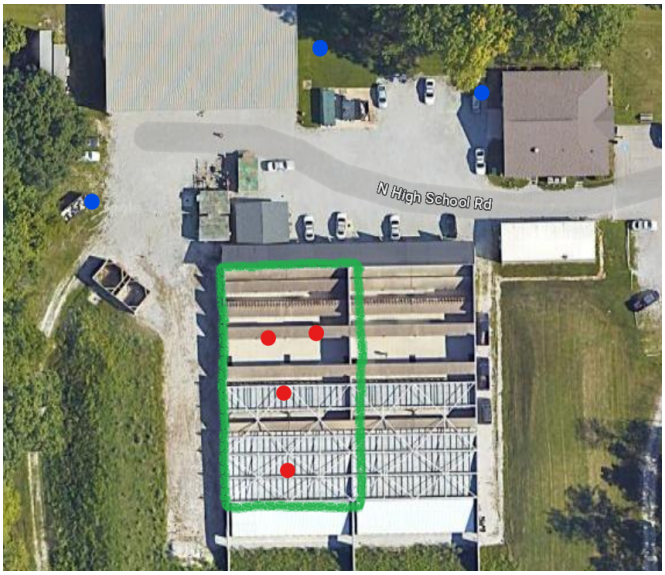


Fig. 7: Unknown Shooter Location Results. All officers were within the green box. The blue points are the locations of the devices, and the red points are the predicted locations of the officers.

We also tested the localization system using the gradient-based TOA identification method outlined above. At this time, our devices do not include GPS based synchronization, so we used recording devices manually synchronized to perform the localization.

There were two situations that our devices recorded:

1) Officers shooting in a semi-enclosed area with exact location unknown. The officers were in the left half of the semi-enclosed area circled in green in Fig. 7. The devices are plotted in blue. The localization system successfully identified all shots as being within this area (red points). It should be noted that in Fig. 7, there are points that overlap completely and cannot be distinguished.
2) A single officer with known location shooting in an open range (Fig. 8). The TOA localization for 11 shots had a mean error of 20.0 meters. The obstruction added systematic error. To account for this systematic error additional detecting devices are required [24]. With the small amount of data here, we observed that the eigenvalue for the error estimation was highly correlated with the error ($r = 0.93$).

VI. Conclusion

When community violence interrupters respond to reported shooting events, research has shown that they can reduce the risk of subsequent retaliatory shootings [25]. Our hypothesis is that low-cost gunshot detection systems such as the one described here can provide real-time information on unreported shots fired, which can facilitate violence interruption in the absence of administrative incident data on which community violence interrupters typically rely. Future research will focus on testing the system outside of experimental conditions to explore the accuracy and robustness of the system when deployed in urban environments. In this setting online model fine-tuning over the cellular network will further improve performance, using labels collected from community intervention workers.

Fig. 8: Known Shooter Location Results. The blue points are the device locations, the green point is the true shooter location, and the red points are the predicted shooter locations.

VII. Acknowledgments

This research was supported by NSF grants SCC-2125319, and ATD-2124313, NIJ grant 2018-75-CX-0003 and AFOSR MURI grant FA9550-22-1-0380. PJB and GM serve on the board of Geolitica, a software analytics company serving law enforcement. We thank Michael Daley and the Indianapolis Metropolitan Police Department for hosting us at the IMPD firearms training range.

Appendix: Hessian of Log Posterior

To make the notation simpler, we define $\mathbf{x_0} = (x_0, y_0)$, $d_i = \sqrt{(x_0 - x_i)^2 + (y_0 - y_i)^2}$, and $\Delta t_i = t_0 - t_i$. We also ignore the constant $\frac{1}{2\sigma^2}$ in front of ℓ because it will be absorbed in the regression model later and σ is not easily found. Additionally, we will not derive expressions for y_0 since they are duplicated for x_0.

$$\frac{\partial d_i}{\partial x_0} = \frac{x_0 - x_i}{d_i} \tag{8}$$

$$\frac{\partial \ell}{\partial d_i} = \frac{-2(d_i + s\Delta t_i)}{s^2} \tag{9}$$

$$\frac{\partial \ell}{\partial x_0} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial d_i}\frac{\partial d_i}{\partial x_0} = -\frac{2}{s^2}\sum_{i=1}^{n}(1 + \frac{\Delta t_i s}{d_i})(x_0 - x_i) \tag{10}$$

$$\frac{\partial^2 \ell}{\partial x_0^2} = -\frac{2}{s^2}\sum_{i=1}^{n}(1 + \frac{s\Delta t_i}{d_i}(1 - \frac{(x_0 - x_i)^2}{d_i^2})) \tag{11}$$

$$\frac{\partial^2 \ell}{\partial x_0 \partial y_0} = \frac{2}{s}\sum_{i=1}^{n}\frac{\Delta t_i (x_0 - x_i)(y_0 - y_i)}{d_i^3} \tag{12}$$

$$\frac{\partial \ell}{\partial t_0} = -2\sum_{i=1}^{n}(\frac{d_i}{s} + \Delta t_i) \tag{13}$$

$$\frac{\partial^2 \ell}{\partial t_0^2} = -2n \tag{14}$$

$$\frac{\partial^2 \ell}{\partial t_0 \partial x_0} = -\frac{2}{s}\sum_{i=1}^{n}\frac{x_0 - x_i}{d_i} \tag{15}$$

REFERENCES

[1] D. Searcey, "What would efforts to defund or disband police departments really mean," *The New York Times*, vol. 4, 2020.

[2] K. Bullock and N. Fielding, "Community crime prevention," in *Handbook of crime prevention and community safety*, pp. 87–108, Routledge, 2017.

[3] B. C. Welsh, G. M. Zimmerman, and S. N. Zane, "The centrality of theory in modern day crime prevention: Developments, challenges, and opportunities," *Justice Quarterly*, vol. 35, no. 1, pp. 139–161, 2018.

[4] C. W. Telep and J. Hibdon, "Community crime prevention in high–crime areas: The seattle neighborhood group hot spots project," 2018.

[5] L. Tompson, S. Johnson, M. Ashby, C. Perkins, and P. Edwards, "Uk open source crime data: accuracy and possibilities for research," *Cartography and geographic information science*, vol. 42, no. 2, pp. 97–111, 2015.

[6] R. Solymosi, K. J. Bowers, and T. Fujiyama, "Crowdsourcing subjective perceptions of neighbourhood disorder: Interpreting bias in open data," *The British Journal of Criminology*, vol. 58, no. 4, pp. 944–967, 2018.

[7] D. C. Folch, S. E. Spielman, and R. Manduca, "Fast food data: Where user-generated content works and where it does not," *Geographical analysis*, vol. 50, no. 2, pp. 125–140, 2018.

[8] A. Morehead, L. Ogden, G. Magee, R. Hosler, B. White, and G. Mohler, "Low cost gunshot detection using deep learning on the raspberry pi," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3038–3044, IEEE, 2019.

[9] L. K. Katsis, A. P. Hill, E. Piña-Covarrubias, P. Prince, A. Rogers, C. Patrick Doncaster, and J. L. Snaddon, "Automated detection of gunshots in tropical forests using convolutional neural networks," *Ecological Indicators*, vol. 141, p. 109128, 2022.

[10] T. Aggarwal, N. Sharma, and N. Aggarwal, "Gunshot detection and classification using a convolution-gru based approach," in *Proceedings of Emerging Trends and Technologies on Intelligent Systems: ETTIS 2022*, pp. 95–107, Springer, 2022.

[11] D. Nieves-Acaron, B. Luchterhand, A. Aravamudan, D. Elliott, S. Wyatt, C. E. Otero, L. D. Otero, A. O. Smith, A. M. Peter, W. Jones, and E. Lam, "Ace: An atak plugin for enhanced acoustic situational awareness at the edge," in *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, pp. 115–120, 2021.

[12] J. Svatos and J. Holub, "Impulse acoustic event detection, classification, and localization system," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023.

[13] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 2021.

[14] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers and distillation through attention," 2021.

[15] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, (New York, NY, USA), p. 1041–1044, Association for Computing Machinery, 2014.

[16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[17] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, B. Lee, and A. Marco, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wirel. Commun. Mob. Comput.*, vol. 2017, jan 2017.

[18] Y. Zhou, "An efficient least-squares trilateration algorithm for mobile robot localization," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3474–3479, 2009.

[19] M. Cobos, J. J. Perez-Solano, S. Felici-Castell, J. Segura, and J. M. Navarro, "Cumulative-sum-based localization of sound events in low-cost wireless acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1792–1802, 2014.

[20] W. C. Chung and D. Ha, "An accurate ultra wideband (uwb) ranging for precision asset location," in *IEEE Conference on Ultra Wideband Systems and Technologies, 2003*, pp. 389–393, 2003.

[21] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 4190–4197, Association for Computational Linguistics, July 2020.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, oct 2019.

[23] G. Yan and J. Tang, "A bayesian approach for localization of acoustic emission source in plate-like structures," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–14, 2015.

[24] R. B. Calhoun, C. Dunson, M. L. Johnson, S. R. Lamkin, W. R. Lewis, R. L. Showen, M. A. Sompel, and L. P. Wollman, "Precision and accuracy of acoustic gunshot location in an urban environment," 2021.

[25] J. Park, F. P. Schoenberg, A. L. Bertozzi, and P. J. Brantingham, "Investigating clustering and violence interruption in gang-related violent crime data using spatial–temporal point processes with covariates," *Journal of the American Statistical Association*, vol. 116, no. 536, pp. 1674–1687, 2021.