Document Title:	Development of Computational Methods for the Audio Analysis of Gunshots
Author(s):	Ryan Lilien
Document Number:	252947
Date Received:	May 2019
Award Number:	2016-DN-BX-0183

Final Research Performance Progress Report - Cover Page

Federal Agency and Organization Element: Department of Justice, Office of Justice Programs

Federal Grant or Other Identifying Number: 2016-DN-BX-0183

Project Title: Development of Computational Methods for the Audio Analysis of Gunshots

PD/PI Name, Title, Contact Info: Ryan Lilien, Chief Scientific Officer, Cadre Research Labs; 420 W Huron St, Suite 204; Chicago, IL 60654

Name of Submitting Official: Ryan Lilien

Submission Date: June 26, 2018

[redacted]

Recipient Organization: Cadre Research Labs, LLC (small business)

Recipient Identifying Number (if any): N/A

Project/Grant Period: Start: 1/1/2017, End: 3/31/2018

Reporting Period End Date: 3/31/2018

Report Term or Frequency: Twice a year

Signature of Submitting Official:

Note that this report aims to follow the NIJ deliverable 10 page double spaced final summary overview format. We created to a 10 page report not including this cover page and not including figures (which appear at the end of this report). We also included a single page appendix that addresses "Implications for Criminal Justice Policy".

1 Project Purpose and Background

Audio analysis of gunshots is based on the observation that the content and quality of gunshot recordings are influenced by firearm and ammunition type, the scene geometry, and the recording device used. Advanced computational techniques can exploit these facts to answer investigative questions. For example, in much the same way that signal processing for human speaker recognition can help reach conclusions regarding the gender, age, identity, or national origin of the speaker, it appears possible that similar methods may be able to answer firearm specific questions from audio data. As more crimes are captured on audio recordings, more examiners will be asked to answer questions regarding firearm-related audio. The completed project aims include the application of advanced computational tools to audio analysis. Our approach uses a fine-grained mathematical representation of the frequency spectrum with a series of advanced machine learning techniques for clustering and pattern recognition. Several specific investigative questions are addressed. The research work was completed by Cadre Research Labs, a scientific computing contract research organization, working in collaboration with experienced firearm examiners (Lucien Haag, Mike Haag, Todd Weller).

2 Project Design

2.1 Gunshot Audio Background

When a firearm is discharged several sounds are produced. These include the muzzle blast, a shock wave (if the bullet is traveling at supersonic speeds), secondary mechanical sounds, the bullet impact, and scattered reflections [1, 8]. The *muzzle blast* refers to the complex acoustic signal arising from the rapid ejection of gas from the firearm muzzle. It is short (on the order of milliseconds) and loud (typically 120-160dB). If the blast is close to the recording device, the volume of the blast may overwhelm the recorder resulting in saturation and spectral information loss. Bullets moving faster than the speed of sound will produce a shock wave propagating outward from the bullet's path. *Secondary mechanical sounds* include sounds related to loading, cocking, firing, and ejection mechanics. In contrast to the muzzle blast, these mechanical sounds may be low volume and difficult to pick-up unless the recording device is close to the discharged firearm. The bullet impact may be detected but is not typically considered when considering the number of shots fired or the identity of each shot. Finally, any of the described sounds can produce reflections when the sound wave bounces off a secondary surface [9]. The most common of these is the *ground reflection* of the muzzle blast.

Several secondary factors effect the recorded audio. First, sounds may arrive at the recording device in a non-chronological order. That is, depending on the scene geometry, the supersonic shock wave may arrive before the muzzle blast [6, 8]. Second, the muzzle blast is highly directional, dependent on the azimuth angle formed between the muzzle direction and the recording device. Both the overall volume and waveform shape vary from a loud and structured sound when pointed towards the recording device (azimuth of 0-degrees) to a quieter less structured waveform when pointed away (azimuth of 180-degrees) (azimuth angle shown in figure 3) [1]. Finally, the audio is susceptible to environmental conditions (temperature, humidity, wind) and scene geometry (absorption, reflection, focusing). The recording is also influenced by the firearm make/model, caliber, and ammunition type. Finally, the recording device itself influences the recording. Each device has a frequency response that describes how efficiently the device captures sound at different frequencies. The audio file format may employ lossy compression methods which introduce sound artifacts[1]. While explicit handling of all these variables is unrealistic, it is important to be aware of their potential influence.

2.2 Project Aims

The completed work was split into two aims. Aim 1 involved the collection of two datasets. Aim 2 involved the development and application of advanced computational tools to answer several questions.

1. Detect gunshots in an audio recording
2. Compute shot-to-shot timings
3. Determine the number of firearms present in a recording and assign shots to firearms
4. Construct a predictive model of the likely class, caliber, and make/model of recorded gunshots

3 Materials and Methods

This proposal has two primary aims. The first aim involves Data Collection and Processing. The second aim involves Audio Analysis. Each aim will be discussed separately. Methods have been abbreviated to conform to the maximum page limit.

[1] Compression techniques are defined as being either lossless, which perfectly preserve the original signal, or lossy where the original signal is not perfectly preserved.

3.1 General Methods and Concepts

(Aim 1) Data Collection: Dataset 1: The first dataset includes test fires for approximately twenty firearms collected using multiple devices at multiple positions relative to the shooter. Firearms were selected based on their commonality, their potential to be used in gun crimes, and the types of sounds they produce. Several makes, calibers, and firearm types are included. Each gunshot was recorded on four different recording devices. These devices include a high-quality Zoom H4N hand-held recorder, an iPhone 7 smartphone, a Samsung Galaxy S7 smartphone, and a Prima Facie BodyCam which are referred to as Zoom, iPhone, Samsung, and Bodycam in the rest of this report. The Zoom device was equipped with a 'fuzzy' wind-brake designed to reduce the effect of wind noise. To increase efficiency, two sets of recording devices were obtained allowing the project team to record data from two locations at the same time. This halved both the amount of ammunition required and the time required to collect the data. We assume that each of the identical devices (*e.g.*, each of the identical iPhones) recorded similarly.

(Aim 1) Data Collection: Dataset 2: The second dataset includes audio files extracted from YouTube and represents the type of 'real-world' data that might be encountered in casework. These files were collected on unknown recording devices, in varying environments, and with a range of background noise.

(Aim 1) Data Processing: Several pre-existing and newly written software tools were used with Datasets 1 and 2 to create a large set of individual processed files. These processed files include short audio clips and spreadsheets of annotations for each file. These spreadsheets include information such as the start and stop time of the individual gunshot as well as the recording device and conditions. This 'meta-data' was used in fitting our statistical models. Each audio file was also converted into a single predefined sampling rate so that they can be compared. The Zoom, Samsung, and Bodycamera record in stereo and the first channel was used for all analysis. The iPhone records in single channel (mono). All data files were converted to 44.1kHz sampling rate and saved as lossless WAV files. This sampling rate was selected as it is extremely common and is the standard for CD recordings. As technology improves we expect a shift to even higher rates. A series of CSV files were created with individual annotations. For example, each experiment recording in Dataset 1 contains six gunshots from a single firearm and the CSV file lists the start and stop times of each gunshot. After splitting each recording such that each individual WAV file contains a single gunshot we created a master CSV which lists details about each file. For example, the firearm, caliber, class, recording position, recording device. This workflow created thousands of files which facilitated our analytic workflow and testing of experiments.

(Aim 1) Gunshot Representation (Features): Every audio signal can be broken down into one or more component sinusoid waveforms. Even the complex audio of human speech is simply a summation of individual sin waves each with different frequencies, start and stop times, and amplitudes (volumes). When analyzing an audio waveform it is common practice to consider the individual frequency components. Several frequency-based approaches were considered to represent the information content of each portion of each audio file. For example, each file is split into a sequence of short (10 or 20 millisecond) segments (aka windows or clips) and a representation was computed for each segment. Representations typically are a summarization of the audio frequencies present in each short segment. These representations serve as the input to secondary analysis methods. Most representations were based around versions of the Mel-Frequency Spectral Coefficients (MFSCs) [2, 3, 12]. The MFSC computes the audio frequencies present in each short time window within an audio file and represents these frequencies as a vector. For example, the MFSC vector might have ten coefficients where each coefficient represents the amount of energy present in the window of a specific frequency band (Figure 1). The closely related Mel-Frequency Cepstral Coefficients (MFCCs) is a variant of the MFSC where an additional step is applied to attempt to uncorrelate the MFSC coefficients. Removing this correlation is beneficial when the signal is band-limited and some of the MFSC terms are non-informative. The theory of this approach is that similar sounds will have similar MFSC (or MFCC) coefficients. This property is what makes MFSCs and MFCCs useful in audio recognition (*e.g.*, speech recognition) and potentially useful in gunshot analysis.

MFSC/MFCCs are often used with delta and delta-delta coefficients. The delta coefficients are the difference between the coefficients at different times. The delta term is $\mathrm{MFSC}_t - \mathrm{MFSC}_{t-1}$ (difference in coefficients at time t and $t-1$) and the delta-delta coefficients are $(\mathrm{MFSC}_t - \mathrm{MFSC}_{t-1}) - (\mathrm{MFSC}_{t-1} - \mathrm{MFSC}_{t-2})$ (the difference in the difference). Although we explored the use of delta and delta-delta terms for each task, we found that explicitly including the straight MFSC terms for multiple consecutive windows performed better. The results below therefore did not utilize the delta or delta-delta terms.

Because muzzle-blasts are extremely short events (typically shorter than 5ms) it is difficult to measure the evolution of frequency components over time. That is, if one split a 5ms muzzle-blast into four different parts (beginning, middle, middle, and end) then each part would represent 1.25ms and would only have 55 sample points (if the original audio was sampled at 44kHz). Therefore, rather than splitting each gunshot into extremely short windows (*e.g.*, 1.25ms) we decided to keep gunshots contained within a single 10ms target window and to compute MFSC/MFCC components for each window. We also modeled the 10ms before the target window and 500ms after the target window. The 10ms before the target was included to capture the background noise before the gunshot occurred. A trained model can therefore take these background frequencies into consideration when evaluating the target 10ms window. In other words, ignore the frequencies in the target window that are also present before the target window. The 500ms after the target window was included to capture the full duration of a gunshot's sound (including echos or delayed response of the recording device). The decision to consider 520ms reflects the fact that our recording devices, sampling rate, environmental conditions, and physical setup captured waveforms that were significantly more noisy than that would be collected under ideal conditions. For example, Maher collects beautiful waveforms using high-quality microphones with dynamic range from 46 dB to 178 dB using a 16-bit recorder at 500kHz sampling [10]. This setup can produce clean waveforms with more than 11-times as many samples as we collect (500kHz vs 44.1kHz). For example, two of our recordings of the same Colt 1911 (.45 Auto) are shown in Figure 2. Both were recorded at 40m and -30 degrees from the shooter. The iPhone recording is typical of those we recorded off cell-phones. The recording shows an approximately 5ms muzzle-blast and is followed by noise which slowly decreases to zero over about 500ms. The Zoom recording of the same exact gunshot also lasts about 5ms but has essentially no noise after the muzzle blast. Because these recordings were collected from the exact same physical event and from the same location we assume the differences are due to the recording devices themselves. Each device has a different frequency response and ability to handle very loud impulses.

(Aim 2) Algorithm Analysis: A number of computational methods were explored for each of the investigated problems. Complete details on each method are beyond the scope of this report. We attempted to provide the high-level intuition behind each method and some specific detail that would be useful for those with experience with these methods. All data analysis and methods development projects are exploratory in that a multitude of methods, approaches, and variants are considered before settling on a method that performs best. In this work, a number of methods were investigated and found to not be useful; many of these methods are not described in this report. A brief description of the primary methods appears next.

Gunshot Analysis (non-neural network based): Several approaches were considered for comparing the audio representations to complete each of the proposed tasks. In general, these methods rely on the assumption that similar sounds will have similar representations (*e.g.*, either MFSC/MFCC or waveform). The following methods are utilized in the experiments below; note that each of these are established methods and a full description of the approach and theory does not fit in this report. In many cases we utilize a *training* set to fit the model and then a *testing* set to evaluate the model (additional detail below). Samples in the training set are **labeled**. For example a label for a 10ms window could be that it contains a gunshot or that it does not contain a gunshot. Another label could be that the gunshot is from a pistol or that it is from a rifle. The samples in the testing set have their labels hidden and it is the job of the algorithm to predict the label using the learned model. The following methods were tested in this work; however, only a few were found to work in this application. **k-Nearest Neighbors (kNN)**: In kNN each sample point in the test set is assigned the label that is most common among its k most similar samples in the training set. **Gaussian Mixture Models (GMM)**: A GMM models the training set by fitting one or more Gaussian (normal) distributions to the labeled training points. These Gaussians represent probability distributions for each label. A sample in the testing set is labeled by determining the most likely label according to the fit probability distribution. **Logistic Regression (LR)**: A logistic regression model can be used to assign a dichotomous label (one where there are only two labels, *e.g.*, gunshot or non-gunshot). LR computes a weighted sum of one or more independent variables (*e.g.*, the components of the MFSC/MFCC or waveform). **Linear Discriminant Analysis (LDA)** [4, 11] is based on the more widely known Principal Component Analysis (PCA) and is a linear method for clustering points based on their assigned labels. A weighted combination of each sample's features are used to map each sample to a location in a lower-dimensional space where similarly labeled samples are close together and differently labeled samples are far apart. For example, these approaches may assign labels based on the magnitude of specific components; for example, the presence of a specific low-frequency band may support the assignment of label X. The model learns the weights and support provided by each component. LDA performed the best for most of the applications listed below.

The problem of determining the number of firearms present in a recording is a variant on the above setup. When determining the number of firearms present (and the assignment of which gunshot corresponds to which firearm) a training set is not explicitly utilized. Instead, the gunshots of a single recording are analyzed as a group where each gunshot is a sample. The goal of the algorithm is to determine the number of clusters (*e.g.*, labels) that best fit the data and then assign these labels to the samples. One way of determining the number of clusters is to consider the quality of the clustering that results when the algorithm is forced to use exactly k labels (for $k = 0 \ldots n$). **k-Means**: k-means performs a clustering to identify k clusters. It randomly assigns k cluster centers and then iteratively optimizes the placement of these cluster centers based on the samples. After the cluster centers have been placed we can evaluate the quality of the clustering by determining the variance within a cluster. The *gapstatistic* [13] allows us to identify which k affords the largest relative improvement compared to $k - 1$. If no

k provides a significant improvement compared to $k - 1$ then the best k is assumed to be one. Ideally, each cluster contains gunshots for a single firearm. In some cases when a recording has gunshots from multiple similar sounding firearms the k-means approach may place gunshots from two firearms in a single cluster.

Gunshot Analysis (neural network based): Neural networks have recently enjoyed success in speech and other pattern recognition tasks. Deep Neural Networks (DNNs) are mathematical models very loosly structured after biological neural networks. Neural networks are a complex topic and the description provided here only scratches the surface; references are provided for the interested reader. In a DNN, an input is presented at the input (or first) layer, information in the form of activations flow through a series of connected layers (via series of mathematical operations), and end up at the output layer. The output layer contains k output nodes each corresponding to one of k different output labels. The internal structure of the DNN (comprising a set of nodes and unknown numerical weights) is setup such that when a member of label X is presented on the input layer that the internal activations induce the strongest output activation on the output node corresponding to label X. Although this sounds like magic, the mathematical operations are relatively simple and mainly consist of multiplications and additions. Training a network consists of learning the internal structure and weights that result in the correct output node being activated for each input. This training is done iteratively, using batches of training data. For each batch of training data, the weights of the hidden units are updated. The updating process uses an optimizer. Each pass through all the training data is called an *epoch*. For each successive epoch, we repeat this process with the same training data, but using the model weights from the end of the previous epoch. In the problem of gunshot detection, the input layer is the representation of the audio window under consideration and the output layer contains two nodes (corresponding to the labels of gunshot and non-gunshot). Countless papers have been written on the topic, an excellent contemporary book on Deep Neural Networks is Goodfellow *et. al.* [5]. One particular model of deep learning which has been utilized in audio processing is the Long Short-Term Memory (LSTM) model [7]. An audio LSTM breaks longer audio signals into a sequence of sounds (which can be thought of as phonemes) which together constitute a word. This concept of breaking a long sound into a sequence of shorter sounds unfortunately is not applicable to the analysis of gunshots given their extremely short duration.

(Aim 2) Training/Testing: Many computational methods within the field of machine learning fit a model to a set of *training* data and then evaluate the performance of the fit model a second, different, set of *testing* data. The work performed here is considered 'supervised' learning in that annotations (or labels) are required for both the training and testing data. In our case, the annotations include the exact time location of the gunshot in the recording as well as the firearm make, class, and caliber. The annotations of the training data are shown to the algorithm and are used while fitting the model. The annotations of the testing data are hidden from the model as the model's task is to predict the labels. We use the labels of the testing data as the 'answer key' against which we evaluate performance. It is therefore important that the training and testing sets contain different data. If the two sets contained the same data then the performance on the testing data would not necessarily be indicative of real-world performance. That is, rather than learn the association between features and labels, the algorithm could over-fit and memorize the label for a given sample. A non-technical example would be testing a classroom of students; if the practice exam and the final exam both contained the same questions, it would not be a great measure of how much the students had learned as it may measure how well they could memorize the practice exam.

In many of the experiments below the data is split into training and testing. We randomly selected all recordings from all firearms from the following four geometries to be the testing set: 20m 180-degrees, 20m 60-degrees, 3m 0-degrees, 12m 75-degrees. The remaining sixteen geometries listed in Figure 3 comprise the training set. This results in an 80%/20% training/testing split. Finally, our datasets had significant class imbalance between the number of windows with gunshots and the number without gunshots. We therefore used a standard machine learning technique of subsampling to only train on a random fraction of the non-gunshots but use all the gunshots.

3.2 Specific Methods for Each Application

In this section we discuss the analytic methods used to address each of the four main questions (gunshot detection, shot-to-shot timings, determine number of firearms present and gunshot assignment, and predicting firearm class, caliber, and make/model.

Gunshot Detection: Several approaches were evaluated for the detection of gunshots within a recording. These included GMMs, kNN, LR, and LDA. We found that LDA with a hard prefilter achieved the best performance. For each target 10ms window we used the following feature set. First, we considered four sequential windows including the 10ms window of interest, the 10ms before the window of interest, and the 500ms after the window of interest (split into two 250ms windows). For each window we computed 35 MFSC coefficients and an energy. The energy (or volume) of a window was computed as the normalized magnitude of the power-spectrum[2]. Overall these four windows span 520ms (just over

[2] The volume of a window can be computed by summing the terms of the power-spectrum. The power-spectrum is the

half a second) and are represented by a 140-dimensional MFSC vector and a 4-dimensional energy vector (Figure 4). The following two methods were evaluated using our datasets and results are reported below:

- **Method 1**: Audio files were split into a sequence of 10ms non-overlapping windows. Each window was evaluated as to its likelihood of being a gunshot. We constructed a method with two stages. The first stage performs a hard filter and looks for three consecutive windows (centered at the 10ms window of interest at time t) where the volume increases (between time $t-1$ and t) and then decreases (from time t to $t+1$). Windows not following this 'short-impulse' profile are marked as non-gunshot. Windows meeting the short-impulse profile are then evaluated by volume. That is, the audio must reach at least 70% of the maximum allowed recording volume. Note that in most cases microphones are designed to record normal human-friendly volumes. Since gunshots are much louder than human-friendly sounds they almost invariably saturate the recording device. The one exception is a firearm recorded at a significant distance. These at-a-distance gunshots often are somewhat nondescript and often sound like generic pops; it is significantly more difficult to identify these poorly recorded quiet gunshots. Windows meeting the short-impulse profile and volume threshold are considered 'candidate' gunshots are continue on to stage two. The second stage only considers windows emerging from the first stage as candidate gunshots. A model was built using Linear Discriminant Analysis (LDA) as described above. The two stage approach is as follows:
  - Stage 1: Prefilter each window t for candidate gunshots by looking for impulses. If the volume increases then decreases over three consecutive windows and if the window at time t has a volume at least 70% of the maximum then the segment is considered a candidate and is evaluated in stage 2.
  - Stage 2: Use LDA to assign a label of gunshot or non-gunshot for each candidate gunshot coming out of stage 1.

- **Method 2**: The second method utilizes a deep Deep Neural Network (DNN). The input representation of each target 10ms window is similar to that described above (140 MFSCs and 4 energy terms). For our DNNs, we considered a variety of network architectures (configuration of intermediate or hidden layers) and found that the best-performing choice has three intermediate layers with 80, 40, 10, and 5 hidden units, respectively. Internal nodes use the rectified linear unit

energy of a waveform split by its different frequency components.

(ReLU) activation function. The output layer has two nodes (with a softmax activation) corresponding to the labels of gunshot and non-gunshot. We utilized the Keras machine learning suite (`http://www.keras.io`) to implement the networks and trained the model using the Adam optimizer. Unlike with the LDA model, we did not utilize impulse filters with the DNN model. The hypothesis is that the neural network model would explicitly learn to detect impulses.

As described for our non-neural network approach, class imbalance was also an issue in our DNN implementations. For our DNN models, we downsampled negatives to achieve a 5:1 ratio of non-gunshots to gunshots. Moreover, we also reweighted the loss function according to this class imbalance during training. That is, there was more of a training penalty incurred when making an error on a gunshot than a non-gunshot.

Shot-to-Shot Timings: A small number of experiments were conducted with semi-automatic or fully-automatic firearms to explore the shot-to-shot timings of each. For these experiments the time of each gunshot was manually identified and the shot-to-shot times computed using simple subtraction.

Determine Number of Firearms Present and Shot Assignment: In contrast to the described MFSC approaches, the determination of the number of firearms present and shot assignment problem directly compared the audio waveform of two gunshots. That is, after the individual gunshots had been identified in the source audio file (either manually or using the gunshot detection methods described above) each identified gunshot was compared to every other identified gunshot to determine how similar they were. The most successful method for this comparison uses the maximum cross-correlation (CCF_{max}) between the two waveforms. This involves first normalizing the magnitude of each signal's waveform and then shifting one signal against the second to identify the position where the two waveforms have the highest correlation. Under our normalization method identical signals will have a CCF_{max} of 2.0 while completely uncorrelated signals will have a CCF_{max} of 0 (completely anticorrelated signals will have a score of -2.0). Our normalization approach involves normalizing the two waveforms being compared at the same time; this allows us to differentiate two waveforms with similar frequency components but different volumes (*e.g.*, two similar firearms located at different distances to the recording device). The drawback of comparing complete waveforms using the CCF is that all frequency components are compared including both those that are informative and those that are distracting for differentiation.

For an audio file with multiple gunshots, we first identify all the gunshots, then for every pair of gunshots we compute their similarity using the CCF as computed over the 520ms windows described above (10ms centered at the muzzle blast, 10ms before, and 500ms after, Figure 4). For k gunshots we have $\frac{k*(k-1)}{2}$ pairwise comparisons. A hierarchical clustering is then performed which predicts two gunshots as having come from the same firearm if they are sufficiently similar. We found that a similarity threshold of 1.7 worked well. That is, two gunshots with a pairwise $\mathrm{CCF}_{\mathrm{max}} > 1.7$ are assigned to the same firearm. If all correlations are less than 1.7 we adjust the max threshold lower and consider the most similar gunshot; however, clustering this way assumes that each firearm is fired at least twice. At the end of this clustering it's possible for the gunshots to all be in the same cluster (indicating that a single firearm was present), in two clusters (indicating that two firearms were present), or in k clusters (indicating that k firearms were present).

Prediction of Model of Class and Caliber: As with the other problems above, we considered several methods including GMMs and kNN; however, the best performance was achieved using LDA on the 140-dimensional MFSC values with four energy terms computed over four sequential windows of 10ms, 10ms, 250ms, and 250ms. A training / testing split was used as described in the results section. We predicted class, caliber, and make/model.

4 Data Results and Analysis

In this section we summarize the experimental results.

(Aim 1A) Data Collection: Dataset 1: Gunshot recordings for twenty firearms under twenty conditions were collected in rural eastern Arizona. The collection site is a flat, treeless, remote environment away from highway, aircraft, and city sounds. Unfortunately, during the recording weekend there was intermittent wind at the shoot site. The wind had minimal effect on the Zoom device but significantly impacted the BodyCam recordings to the extent that even after processing the BodyCam recordings were deemed unworthy of analysis. The selected firearms are listed in Table 1. Test set 1 contains three subsets. The first subset includes recordings of the twenty firearms collected at the twenty locations shown in Figure 3. These recording locations vary from a distance of 0m to 40m and an azimuth angle of -30 to 180 degrees. Six gunshots were recorded for each firearm at each location. Note that there were a few combinations of firearm and recording location that were not collected. This either occurred because of ammunition limitations, time limitations, or failure of the recording device that was only noticed after the recording session. The second subset includes recordings with two or three different firearms within the same audio file (Figure 5 and Tables 6 and 7). The third and final subset includes special conditions including supersonic, rapidly fired semi-automatic, and fully automatic firearms. Approximately 13,400 gunshot recordings were collected across all conditions and devices.

A second dataset (Dataset 2) containing 'real-world' test fires extracted from YouTube videos was also obtained. Videos were identified containing the keywords shown in Table 2. Slight variants on each keyword were allowed. The recording environment (indoors or outdoors) and relative recording position (near or far) were noted for each recording. The recording quality of these videos was generally much lower than that of Dataset 1. The collected data was used as described below.

(Aim 1B) Data Processing: All audio files of Datasets 1 and 2 were processed as described above. The initial video files were split by experiment, the audio was extracted for each experiment, and then converted to single audio channel at 44.1kHz and saved in the lossless WAV format. For experiments where we needed individual gunshots for training, we manually went through and split each WAV file into individual files each approximately 1-2 seconds long and containing a single gunshot.

(Aim 1C) Website: A simple website was created to house Dataset 1 and the associated metadata files. The site at `www.CadreForensics.com/audio` provides a small amount of background information on the data, the NIJ disclaimer, and allows visitors to download the data. Users are required to provide contact information so we can keep track of the number and distribution of those using the data.

(Aim 2A) Gunshot Detection Method 1: Method 1 utilizes an energy-based pre-filter to look for loud impulses and an LDA model to differentiate loud impulse gunshots from loud impulse non-gunshots. The combined results for the Zoom, iPhone, and Samsung recordings are presented in Table 3. An 80%/20% training/testing split was used. Note that the use of the loud impulse filter significantly cuts down on the number of false positives (*i.e.*, falsely detected gunshots). However, this impulse detection filter also reduces the number of detected gunshots (*i.e.*, the true positive rate). The YouTube rows show the performance of the model trained on the Zoom, iPhone, and Samsung data but evaluated on the YouTube recordings. The YouTube data has a larger number of false positives, likely due to the fact that most of these recordings are taken close to the shooter and are very loud. The impulse filter greatly improves the performance on the YouTube dataset. Approximately 84% of the gunshots in the YouTube set are detected with this model.

(Aim 2A) Gunshot Detection Method 2: Method 2 utilized a deep neural network based off MFSC and energy terms. We considered datasets for Zoom, Samsung, and iPhone devices as well as a dataset drawn from YouTube audio. We trained our DNN model with a batch size of 10 examples over 100 training epochs. Several models were trained. First, we trained one model for each device using the same 80%/20% training/testing split as used in method 1. As in Method 1, we also considered training on all three devices together (referred to as the 'combined' model). Results are presented for the three devices, the combined model, and the YouTube dataset in Table 4. The results on the Zoom, iPhone, and Samsung data are arguably better than method 1 above. Method 2 does not use a impulse filter yet the detection performance is comparable to Method 1 when a filter is used. This suggests that the DNN model has learned to detect impulses.

Overall, all of our models achieve fair to good accuracy. A few general conclusions can be made. For method 1, the impulse filters do a good job of reducing false positives but have the unwanted effect of reducing the true positive rate. The performance is also effected by the fact that gunshots are a rare occurrence in our datasets which induces class imbalance and a potentially biased model. The neural network approach, (Method 2) agnostically learns this imbalance and thus the false negative rate is likely to be higher in the noisier datasets. When considering results broken out by firearm for each test set (not shown), the false positive identifications are fairly evenly spread across the firearms present in the test set. In contrast, false negative identifications clearly appear in test sets for lower caliber firearms which tended to have far lower volumes. This effect is more pronounced for the Zoom device, which tended to attenuate the recordings.

The generalized performance on the YouTube set is slightly better in Method 2. The YouTube set contained a lot of firearms recorded up-close and significantly more background noise than our core sets. It is interesting that Method 2 was better able to handle these differences. Overall, Method 2 may be a better approach than Method 1; however, both methods may have difficulty generalizing to the full range of recording devices and environments encountered in actual casework. The initial performance is promising and more research is required.

(Aim 2A) Gunshot Timing: Eight datasets were used to determine shot-to-shot timings. The experiments, number of shots, average shot-to-shot timing, and variance of shot-to-shot timing is reported in Table 5. Although this was a small part of our project the results were as expected in that quickest and most consistent shot-to-shot times were found on the fully-automatic Colt M16 rifle. Analyzed recordings were made on the Zoom handheld device at 20m and -30 degrees although any position would have produced the same results.

(Aim 2B) Number of Firearms Present and Annotation: Forty six datasets with one, two, or three firearms were analyzed. Each dataset was recorded with the Zoom handheld device at one of six different geometries (Figure 5). Geometry A refers to a single firearm. Geometries B and F have multiple different firearms at the same physical position. Geometry C is unique in that the recording device is down range and thus the recorded volumes are likely very loud compared with the other geometries where the recording device is 20 meters behind the shooters. Geometry D is a moving scenario where the second shooter takes a series of steps between each shot carefully moving from 10 meters to 20 meters from the first shooter. Geometry E has three firearms each at a different position. Geometry G has two firearms at one position and a third firearm at a second position.

Overall the algorithm performs extremely well under these recording scenarios. Table 6 shows the results of the single firearm experiments. Only two mistakes are made on this set. The Bolt Action 22 is mistakenly heard as two different types of sounds likely due to the variability in the audio produced by the bolt action mechanism itself. A second error is one of the M&P 40s where the algorithm felt two of the shots sounded different enough to predict a second firearm. For the two and three firearm experiments (Table 7) the results were equally strong. The algorithm correctly identified two firearms for all firearm sets at geometries B and D. Mistakes were only made with geometry C. This is the setup where the recording device is down range and therefore we expect that the gunshots saturate the recordings and therefore all sound the same, thereby resulting in only a single firearm detected. Note that these three experiments (012G, 013G, and 014G) would be extremely difficult under any geometry as 012G and 014G each include two of the same make/model firearm and 013G includes two highly similar 9mm firearms. Perhaps most surprising and impressive is the ability of the algorithm to detect two firearms for experiments 010G and 011G as these experiments have two identical firearms are the same position. What might be detected here are effects of the two different shooters. The two shooters have slightly different height, stance, and apparel. After shooting the first shooter asymmetrically stepped aside. Since we were recording behind the shooters it's possible these effects influenced the recording. The algorithm also correctly identified three firearms for all the three firearm experiments.

We note that these results are not perfect and that this approach will not yet generalize to success in random crime scene videos. In none of the experiments did we move the recording device. Movement is likely in video grabbed from a cell-phone. Although geometry D had one moving shooter they moved slowly and over a limited range. During an actual crime it's likely that movement is wildly erratic. The best results (and those reported here) came from the handheld Zoom device. While this is not a high-frequency sampling device with fancy microphone it did collect significantly cleaner recordings than the cell-phones in our study. Overall these results are extremely promising but in its current form the algorithm is likely to have difficulty in real-world scenarios.

(Aim 2C) Class, Caliber, and Make/Model Prediction: The goals of class, caliber, and make/model prediction were quite ambitious and we only expected modest recognition performance. We used the same 140-dimensional MFSC vector with 4-dimensional energy vector to describe each gunshot. LDA was used to cluster based on class, caliber, or make/model. The same training and testing splits as above were used to test the model. The data has three class labels: revolver, pistol, and rifle. The data has ten calibers (.22LR, .380 Auto, .38 SPL, .357 MAG, 9mm Luger, .40 S&W, .45 Auto, .223R/5.56, 7.62x39, .308W/7.62). The data has 18 make/models (firearms 1-20 in Table 1). Therefore, random guessing would have an accuracy of 33% for class prediction, 10% for caliber prediction, and 5.5% for make/model prediction. The results listed by firearm class subset (revolver, pistol, or rifle) as collected on the Zoom device are shown in Table 8). All results are significantly better than random guessing.

Class prediction is the easiest of the three objectives and the results are fairly decent on both the training and testing sets. This suggests that the model is generalizable. For example, 97% of the pistols in the training set are correctly predicted to be pistols and 88% of the pistols in the testing set are correctly predicted to be pistols. The results for caliber and make prediction show signs of over-fitting in that the training performance is significantly better than testing performance. Despite these signs there are still several promising results. For example, the model is able to predict the caliber of pistols with 69% accuracy on the training set and 43% on the testing set. The make/model of a pistol is only predicted with 25% accuracy on the testing set; however, this is still significantly better than the 5.5% expected by random guessing. Overall, the performance is very promising.

5 Scholarly Products Produced

The primary product of the proposed research is the presentation of our results and progress. We plan on writing up our results for publication. We aim to submit a journal version of our grant final report to either the AFTE journal or the Journal of Forensic Science. We also produced what is to our knowledge the largest publicly available set of gunshot recordings. The data is freely available from our website. The site requires users to create an account so that we can keep track of the number of individuals who have downloaded the data. The website is available at: `www.CadreForensics.com/audio`

6 Summary

We successfully completed the proposed aims during the project period. Towards Aim 1, we collected a large set of gunshot audio recordings for approximately twenty different firearms. Recordings were collected on several different devices. We also assembled a large set of ‘real-world’ gunshot recordings from a popular social-media sharing site. Because we only have rights to our own data, we made all our audio data publicly available through a simple download website. Towards Aim 2, we developed and evaluated several algorithmic approaches for the analysis of gunshot audio data. Through the year we continued collaboration with academic and government colleagues. We are preparing our results for publication in a journal article.

We found that we were fairly successful in being able to identify gunshots, extremely successful in being able to determine the number of different firearms present, and moderately successful in the challenge of recognizing firearm class, caliber, make/model. During this investigation we evaluated the performance of a number of contemporary machine learning algorithms. These algorithms were selected based on their performance on other pattern recognition tasks. They include linear and non-linear methods based on both frequency-based and time-based representations. Overall, during this short project we successfully demonstrated that useful information is indeed contained within gunshot audio recordings; however, the information is difficult to extract from lower-quality recordings such as cell-phones and body cameras. High-quality recordings such as custom microphone arrays and high-sampling rate devices may provide the highest quality data and therefore the best chance of success towards information extraction. The results of our preliminary investigations are just a small step towards the development of a useful tool for law enforcement and forensic examiners. The results are extremely promising and we’re optimistic on the future of this application. With further development these methods may become useful to practitioners.

Appendix

Implications for Criminal Justice Policy and Practice

Our primary impact has been supporting the hypothesis that useful information is contained within the audio recordings of gunshots. Our work product includes what is to our knowledge the largest collection of publicly available gunshot audio data. These recordings are now available on our website. Our results and the availability of our data may serve as a next-step towards a fully automated gunshot analysis system. By publishing our results and making our data available, we are enabling other researchers to build on our progress.

Our work directly addresses several aims of the NIJ's Applied Research and Development in Forensic Science for Criminal Justice Purposes program. Specifically we have developed measurement and analytic techniques, grounded in mathematical science that are able to provide accurate quantitative sample comparison. This work benefits the criminal justice system and their ability to extract firearm related information from evidence. For example, our work is a step towards being able to determine the number of firearms present in a recording and providing the critical piece of information regarding which firearm fired first. In this project, we collaborated with three firearms examiners (Todd Weller, Lucien Haag, and Michael Haag). Todd and Luke are currently private practitioners and Mike is an examiner at the Albuquerque police department.

References

[1] S. Beck, H. Nakasone, and K. Marr. Variations in recorded acoustic gunshot waveforms generated by small firearms. *J. Acoustic Society of America*, 4:1748–59, 2011.

[2] J. Bridle and M. Brown. An experimental automatic word-recognition system. *Joint Speech Research Unit*, Report 1003, Ruislip England, 1974.

[3] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28:357–66, 1980.

[4] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–88, 1936.

[5] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 1st edition edition, 2016.

[6] L. Haag. The sound of bullets. *AFTE Journal*, 34:31–42, 2002.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–80, 1997.

[8] R. Maher. Acoustical characterization of gunshots. *Proc. IEEE SAFE 2007: Workshop on Signal Processing Applications for Public Security and Forensics*, pages 109–13, 2007.

[9] R. Maher. Acoustical modeling of gunshots including directional information and reflections. *Proc. 131st Audio Engineering Society Convention*, page Paper 8494, 2011.

[10] R. Maher and T. Routh. Gunshot acoustics: pistol vs. revolver. *Proc. Audio Engineering Society Conference on Audio Forensics*, 2017.

[11] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience, 2004.

[12] P. Mermelstein. Distance measures for speech recognition, psychological, and instrumental. *Pattern Recognition and Artificial Intelligence*, pages 374–388, 1976.

[13] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Socity*, 63:411–23, 2001.

Figures and Tables

1) High Standard Sport King (SpKing22) [.22LR, Pistol] This semi-automatic pistol can be fired very rapidly giving the impression that the shots are from a full automatic firearm

2) S&W 34-1 (S&W22) [.22LR, Revolver] Different in design and operation from semi-automatic pistols but can discharge the same cartridges as the High Standard pistol

3) Ruger 10/22] (Ruger22) [.22LR, Carbine] Common rifle can discharge subsonic and supersonic .22LR ammunition as well as the same ammunition as fired in the previous firearms

4) Remington 33 Bolt-Action Rifle (BoltAction22) [.22LR, Rifle] A bolt action rifle, firing the same ammunition as the Ruger 10/22, lacks the sounds produced by the cycling of the Ruger's semi-automatic mechanism

5) Lorcin L380 (Lorcin380) [.380 Auto, Pistol] Common semi-automatic handguns often involved in gang and drive-by shootings, they are of straight blowback design and have very few moving parts

6) S&W 10-8 (S&W38) [.38SPL, Revolver] Other than the possible sound of cocking the hammer of a revolver, they only produce the sound of the discharge, can use subsonic and supersonic ammunition

7) Ruger Blackhawk (Ruger357) [.357 MAG, Revolver] Similar to .38SPL except that .357 Magnum revolvers are higher powered, producing very loud discharges and launching bullets at supersonic velocities

8) Glock 19 (Glock9A) [9mm Luger, Pistol] Common handguns, particularly with law enforcement, some makes have different systems of operation which could produce unique sounds

9) Glock 19 (Glock9B) [9mm Luger, Pistol] (Duplicate of Firearm 8)

10) Sig P225 (Sig9) [9mm Luger, Pistol] Common handguns, particularly with law enforcement, some makes have different systems of operation which could produce unique sounds

11) M&P 40 (M&P40A) [.40 S&W, Pistol] Common handguns, particularly with law enforcement

12) M&P 40 (M&P40B) [.40 S&W, Pistol] (Duplicate of Firearm 11)

13) HK USP Compact (HK40) [.40 S&W, Pistol] Common handguns, particularly with law enforcement

14) Glock 21 (Glock45) [.45 Auto, Pistol] Large caliber handguns popular with law enforcement, each has a different system of operation which could produce unique sound

15) Colt 1911 A1 (Colt45) [.45 Auto, Pistol] Large caliber handguns popular with law enforcement, each has a different system of operation which could produce unique sound

16) Kimber Tactical Custom (Kimber45) [.45 Auto, Pistol] Large caliber handguns popular with law enforcement, each has a different system of operation which could produce unique sound

17) M16A1 AR15 (M16223)[.223R/5.56, Rifle] Bolt-action very common semi-automatic with law enforcement and civilian shooters. Have been used in some high profile crimes.

18) WASR 10/63 AK47 (WASR762) [7.62x39mm, Carbine] Numerous semi-automatic versions of the AK47 in circulation, would allow the cyclic rate and shot-to-shot time intervals of full automatic fire

19) Winchester M14 (Win308) [.308W/7.62, Rifle] Most common caliber for military and police snipers

20) Remington 700 (Remington308) [.308W/7.62, Rifle] Another common caliber for military snipers and long range shooters, produces a very loud discharge

21) Rock River LAR-15 (RockRiver300) [.300 Blackout, Rifle]

22) Russian SKS (SKS762) [7.62x39mm, Rifle]

23) PWS MK107 Mod 1 (PWS762) [7.62x39mm, Pistol]

Table 1: **Firearm List.** Firearms 1 through 20 were used in most experiments. Firearms 21, 22, and 23 were used in select experiments. Name in parentheses is the short name used in other result tables.

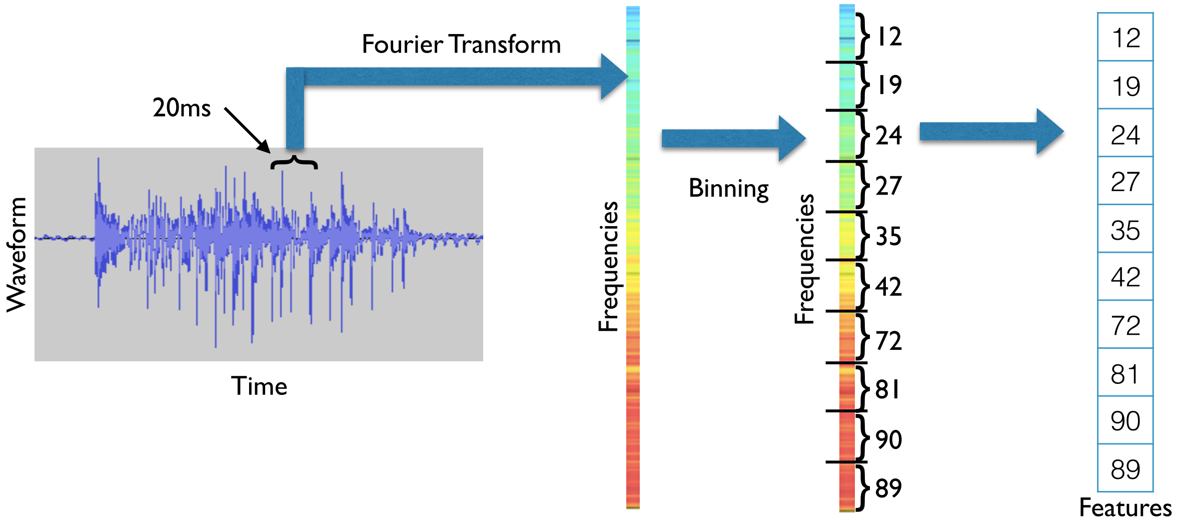

Figure 1: **Features.** Cartoon example of computing an MFSC feature vector for an audio waveform. The frequency spectrum of each window is computed via the Fourier Transform. The power in each frequency band is quantified and converted into a vector (array). In this simplified example, the selected window is represented by the 10 numbers in the vector at the right. Note that this is an extremely simplified example, frequency bands often overlap and have nonlinear spacing and weights.

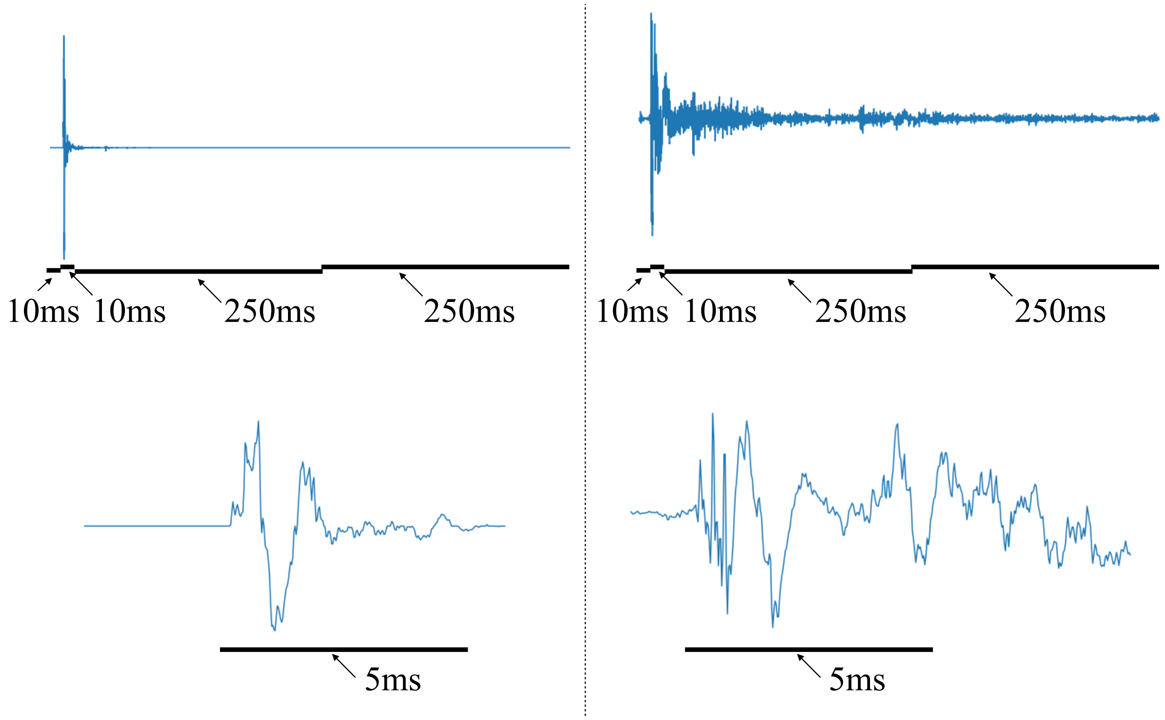

Figure 2: **Colt 1911 Gunshots.** The same Colt 1911 (.45 Auto) gunshot recorded at 40 meters and -30 degrees from the shooter on the Zoom (left) and iPhone (right). The Zoom recording is relatively clean with little extent beyond the muzzle-blast. The iPhone recording is more noisy with echos and reflections lasting 500ms past the muzzle-blast. Close-ups (bottom row) show the 5ms centered on the muzzle-blast.

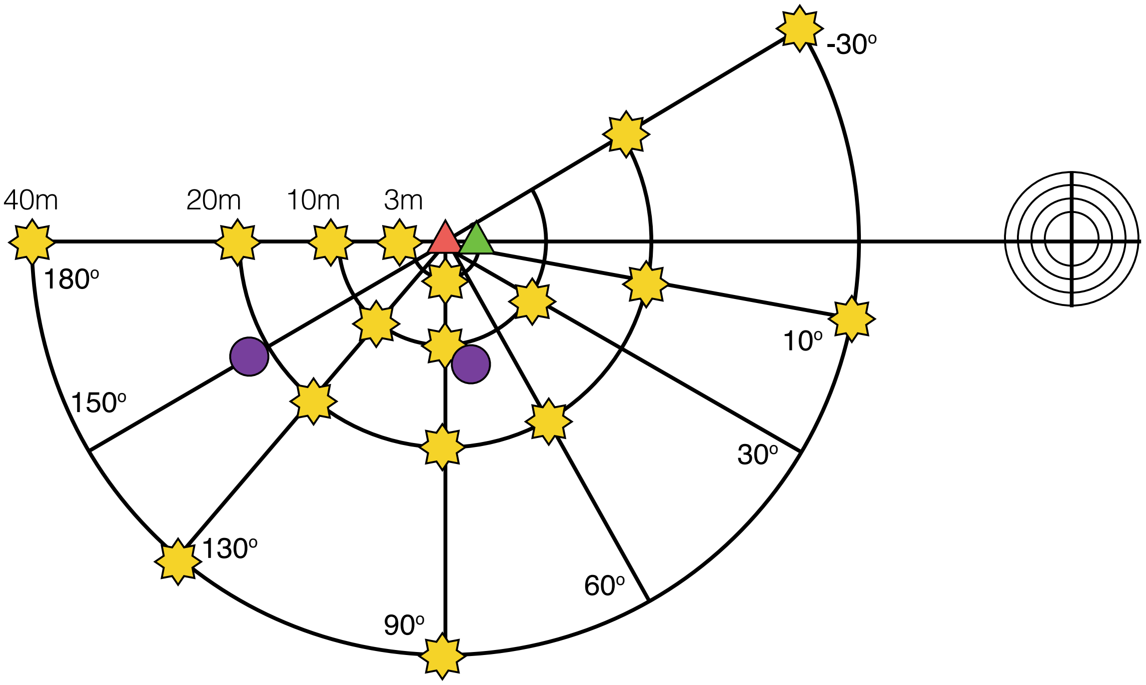

Figure 3: **Collection Geometries for Single Firearm Experiments.** Recordings were collected at twenty different locations (distance and azimuth angle shown). The shooter was located in the center (red triangle). All shots were towards the target (shown at right). All four recording devices were used at each primary location (yellow stars) and secondary location (purple circles). Only the body camera was recorded on the shooter (red triangle). The Zoom, iPhone, and Samsung were recorded on the ground at 3 meters and zero-degrees (green triangle).

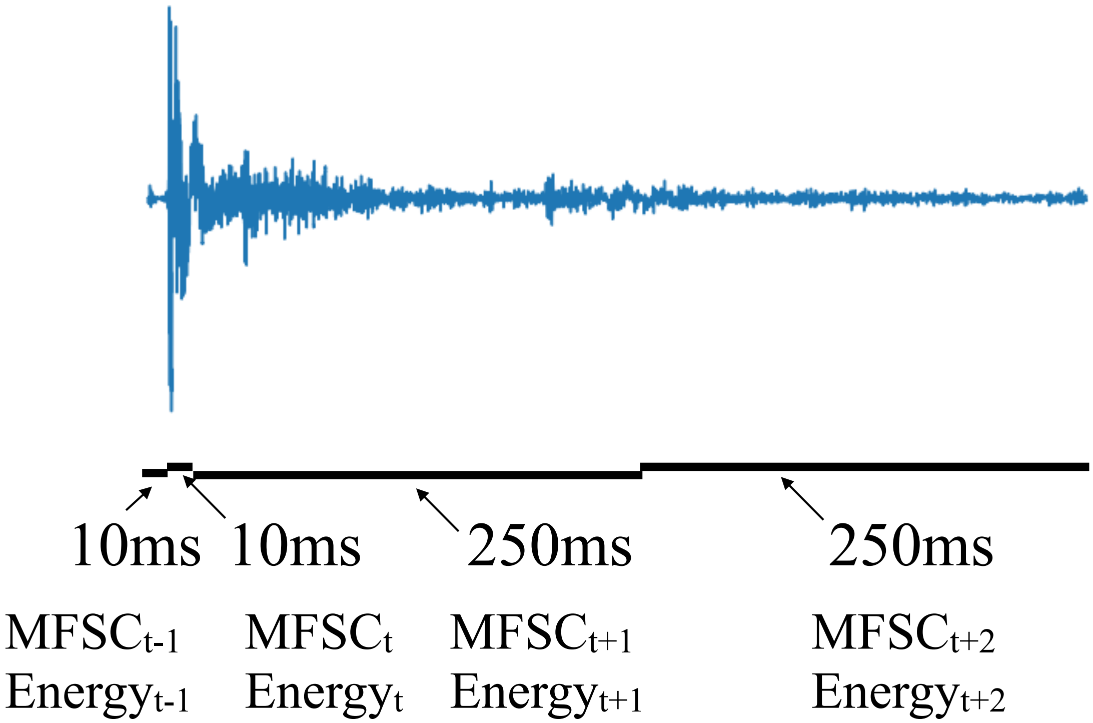

Figure 4: **MFSC and Energy Representation.** The window at time t is represented with four MFSC terms (35 in each of four windows) and four Energy terms (1 from each of the four corresponding windows).

Firearm Type	Number of Audio Files	Approx Number of Gunshots
High Standard Sport King 22	22	132
Ruger Blackhawk 357 Revolver	150	900
Smith & Wesson 22 Revolver	80	480
Ruger 10/22 Carbine	137	822
Remington 22LR	130	780
Lorcin 380	21	126
Smith & Wesson 38 Special Revolver	93	558
Glock 19 9mm	132	792
Sig Sauer P225 9mm	50	300
Smith & Wesson M&P 40	135	810
Heckler & Koch USP Compact 40	49	294
Glock 21 45 Auto	150	900
Colt 1911 45 Auto	115	690
Kimber Tactical Custom 45 Auto	66	396
Colt M16 A1 223R 5.56	75	450
Romarm WASR 10/63 308	50	300
Winchester M14 308	45	270
Remington 700 308	63	378

Table 2: **Dataset 2 (YouTube).** Where “Smith & Wesson” is listed, we also searched for “S&W”, similarly we accepted “H&K” for Heckler & Koch

Train	Test	Filters	TNR	FPR	FNR	TPR
Zoom	Zoom	No	99.98% (37413)	0.02% (6)	0.50% (2)	99.50% (399)
Zoom	Zoom	Yes	100.00% (37419)	0.00% (0)	34.66% (139)	65.34% (262)
iPhone	iPhone	No	99.96% (41723)	0.04% (18)	1.81% (8)	98.19% (435)
iPhone	iPhone	Yes	100.00% (41739)	0.00% (2)	4.97% (22)	95.03% (421)
Samsung	Samsung	No	99.93% (28682)	0.07% (7)	1.99% (6)	98.01% (295)
Samsung	Samsung	Yes	100.00% (28702)	0.00% (1)	10.96% (33)	89.04% (268)
Combined	Combined	No	99.37% (99262)	0.63% (632)	0.57% (11)	99.43% (1928)
Combined	Combined	Yes	99.97% (99860)	0.03% (34)	12.22% (237)	87.78% (1702)
Combined	YouTube	No	78.85% (2159)	21.15% (579)	1.51% (12)	98.49% (782)
Combined	YouTube	Yes	99.05% (2712)	0.95% (26)	15.99% (127)	84.01% (667)

Table 3: **Gunshot Detection (LDA Model).** Table lists the detection results for the testing sets both with and without the use of the impulse detection filters. The first six rows show the results for training and testing on the same recording device. The final four rows train on a Combined set including data from the Zoom, iPhone, and Samsung recordings. The final two rows show the results of the combined model when tested on the YouTube dataset. In all tables: TNR: True Negative Rate, FPR: False Positive Rate, FNR: False Negative Rate, TPR: True Positive Rate. For each result the percentage and absolute number of windows are listed. For example, the model trained and tested on the Combined datasets without impulse filters correctly classifies 99.37% of the non-gunshot windows as non-gunshots and 99.43% of the gunshot windows as gunshots. The FPR and FNR columns indicate non-gunshots which are thought to be gunshots and gunshots thought to be non-gunshots respectively.

Train	**Test**	**Pos**	**Neg**	**TNR**	**FPR**	**FNR**	**TPR**
Zoom	Zoom	407	63125	99.96% (63099)	0.04% (26)	5.90% (24)	94.10% (383)
iPhone	iPhone	449	70008	99.82% (69885)	0.18% (123)	0.22% (1)	99.78% (448)
Samsung	Samsung	307	47587	100.00% (47586)	0.00% (1)	1.95% (6)	98.05% (301)
Combined	Combined	1163	180720	99.83% (180407)	0.17% (313)	0.69% (8)	99.31% (1155)
Combined	YouTube	793	24842	96.70% (24022)	3.30% (820)	2.77% (22)	97.23% (771)

Table 4: **Gunshot Detection (Neural Network Model).** Overall results for each of the neural network models on the testing sets. The first three results rows (Zoom, iPhone, Samsung) were trained on data from the specified device and tested using different recordings from the same device. In the fourth and fifth rows the term ‘Combined’ refers to a combined set of Zoom, iPhone, and Samsung recordings. The fourth row, is trained on all three devices and tested on different recordings from all three devices. Finally, the last row was trained on all three devices and tested on the YouTube dataset. Pos, Neg: The number of positive and negative examples in the set.

Firearm	**Pull Type**	**Number Shots**	**Mean (ms)**	**Variance**
High Standard Sport King .22LR	‘Rattle Finger’ Pistol	11	125	9
High Standard Sport King .22LR	Semi-Auto Pistol	5	238	6
Glock 19 9mm	Semi-Auto Pistol	9	213	23
Sig Sauer P225 9mm	Semi-Auto Pistol	8	249	88
Smith & Wesson M&P .40 S&W	Semi-Auto Pistol	10	200	58
Colt M16 A1 .223R/5.56	Full-Auto Rifle	9	70	4
Smith & Wesson .38 SPL	Semi-Auto Revolver	6	263	12
PWS Pistol 7.62x39mm	Semi-Auto Pistol	10	178	14

Table 5: **Shot-to-Shot Timings.** Table lists the firearms used and the shot-to-shot timings. ‘Rattle-finger’ is a method where a stiffened trigger finger is rattled between the trigger and trigger guard to simulate full-auto fire; the method can only be used on certain firearms. Compare the rattle-finger timings to those of normal operation for the Sport King. As expected, the fully-automatic Colt M16 has the smallest mean shot-to-shot timing with the smallest variance.

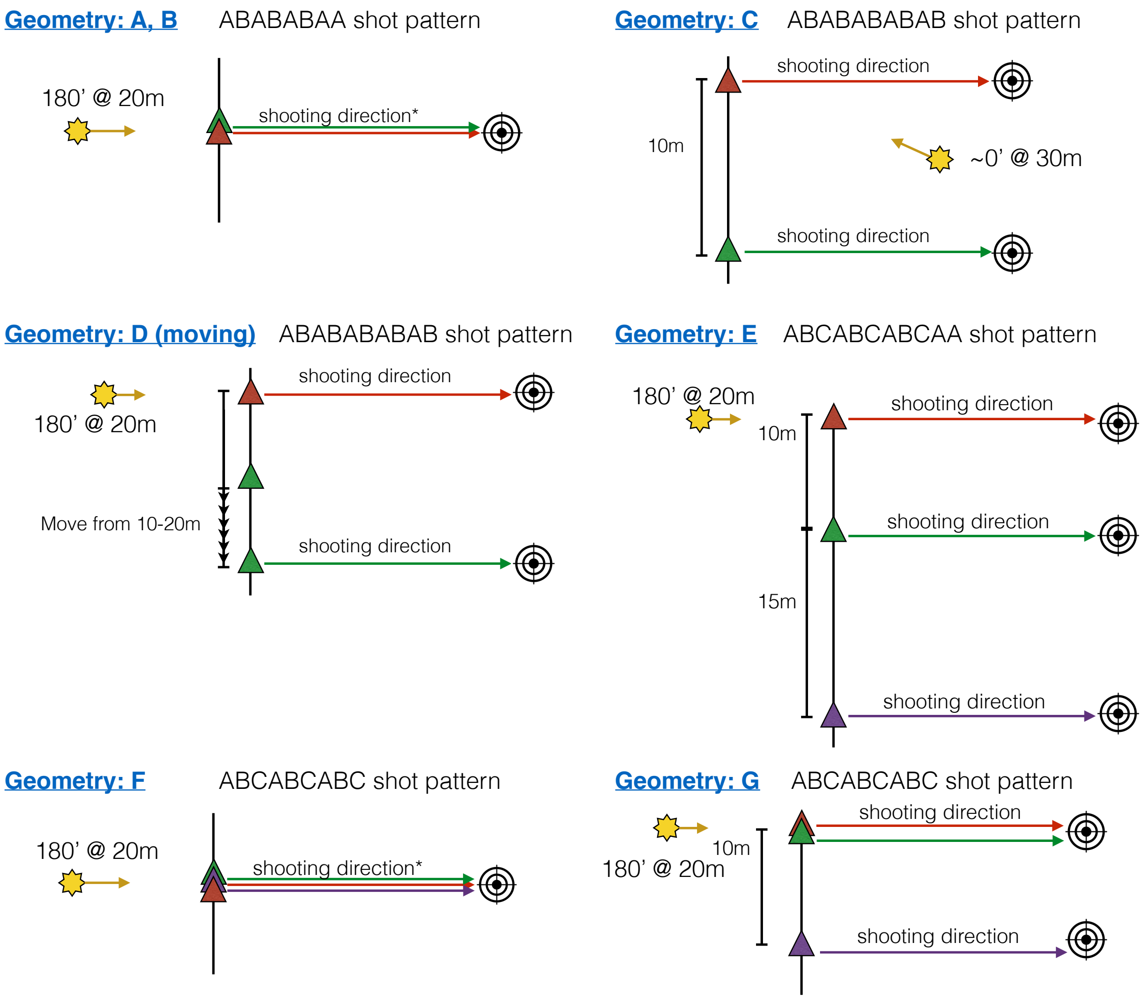

Figure 5: **Collection Geometries for Dual Firearm Experiments.** Recordings were collected at two different locations for each geometry (only one geometry is shown here). Yellow stars indicate the position of the recording devices and the arrow indicates the direction the microphones were pointed. Geometries are referenced in Tables 6 and 7. Recording devices were at 180 degrees and 20 meters for all geometries except B where the recording device was in front of the shooter (0 degrees 30 meters). The position of firearm 1 (red), firearm 2 (green), and firearm 3 (purple) are shown. In geometries B and F all firearms are at the same position. Geometry D presents a moving scenario where firearms 2 moves from a distance of 10 meters from the shooter to 20 meters.

Exp.	**Geometry**	**FA1**	**Detected FA**	**Correct FA Assignment**
081A	A	SpKing22	1	6/6 100%
082A	A	S&W22	1	5/5 100%
083A	A	Ruger22	1	6/6 100%
084A	A	BoltAction22	**2**	4/6 66%
085A	A	Lorcin380	1	6/6 100%
086A	A	S&W38	1	6/6 100%
087A	A	Ruger357	1	6/6 100%
088A	A	Glock9A	1	6/6 100%
089A	A	Glock9B	1	6/6 100%
090A	A	Sig9	1	6/6 100%
091A	A	M&P40A	**2**	4/6 66%
092A	A	M&P40B	1	6/6 100%
093A	A	HK40	1	6/6 100%
094A	A	Glock45	1	6/6 100%
095A	A	Colt45	1	6/6 100%
096A	A	Kimber45	1	6/6 100%
097A	A	M16223	1	6/6 100%
098A	A	WASR762	1	6/6 100%
099A	A	Win308	1	6/6 100%
100A	A	Remington308	1	6/6 100%

Table 6: **Number of Firearm Detection and Shot Assignments (Single Firearm).** Table lists the experiment identifier, the geometry of the shooter (see Figure 5), and the firearm. Detected FA is the number of firearms detected by our algorithm and Correct FA Assignment is the number of gunshots that are correctly assigned to each firearm. For example in experiment 81A there was one firearm detected and all six shots were attributed to this firearm. Results get more interesting in Table 7.

Exp.	Geometry	FA1	FA2	Detected FA	Correct FA Assignment
001G	B	SpKing22	Ruger357	2	8/8 100%
002G	B	SpKing22	Glock9B	2	8/8 100%
003G	B	Glock9A	BoltAction22	2	7/8 88%
004G	B	Glock9A	Sig9	2	8/8 100%
005G	B	M&P40A	HK40	2	8/8 100%
006G	B	Glock9A	M&P40A	2	8/8 100%
007G	B	SpKing22	WASR762	2	8/8 100%
008G	B	Glock9B	M16223	2	8/8 100%
009G	B	Colt45	M16223	2	8/8 100%
010G	B	Glock9A	Glock9B	2	8/8 100%
011G	B	M&P40A	M&P40B	2	8/8 100%
019G	B	M16223	RockRiver300	2	8/8 100%
012G	C	Glock9A	Glock9B	**1**	5/10 50%
013G	C	Glock9A	Sig9	**1**	5/10 50%
014G	C	M&P40A	M&P40B	**1**	5/10 50%
015G	D	WASR762	SpKing22	2	10/10 100%
016G	D	Glock9A	M&P40A	2	10/10 100%
017G	D	Ruger357	Glock9B	2	10/10 100%
018G	D	Glock9A	Glock9B	2	10/10 100%

Exp.	Geometry	FA1	FA2	FA3	Detected FA	Correct FA Assignment
001I	E	SpKing22	Glock9A	Glock45	3	11/11 100%
002I	E	Glock9A	Glock45	RockRiver300	3	11/11 100%
003I	E	S&W22	M&P40A	Colt45	3	11/11 100%
004I	E	Glock9A	Glock9B	WASR762	3	11/11 100%
005I	F	SKS762	WASR762	PWS762	3	9/9 100%
006I	F	SKS762	Glock9A	PWS762	3	9/9 100%
007I	G	SKS762	WASR762	PWS762	3	9/9 100%

Table 7: **Number of Firearm Detection and Shot Assignments (Two Firearms, top) (Three Firearms, bottom).** Table lists the experiment identifier, the geometry of the shooters (see Figure 5), and the firearms. Detected FA is the number of firearms detected by our algorithm and Correct FA Assignment is the number of gunshots that are correctly assigned to each firearm. For example in experiment 001G there were two firearms detected and all eight shots are correctly assigned to their respective firearms.

Prediction	Subset	Train Accuracy	Test Accuracy
Class	Revolver	0.74	0.56
Class	Pistol	0.97	0.88
Class	Rifle	0.90	0.81
Caliber	Revolver	0.86	0.51
Caliber	Pistol	0.69	0.43
Caliber	Rifle	0.88	0.58
Make-Model	Revolver	0.85	0.53
Make-Model	Pistol	0.62	0.25
Make-Model	Rifle	0.76	0.48

Table 8: **Class, Caliber, and Make Predictions.** On this challenging task the performance is significantly better than random chance. As expected, the performance on the training set is better than on the testing set. Class prediction is easier than caliber or make prediction and the results are better. Results are shown for the Zoom recordings.