Audio Engineering Society

Conference Paper

Presented at the Conference on
Audio Forensics
2019 June 18 – 20, Porto, Portugal

Improved gunshot classification by using artificial data

Christian Busse[1,2,3], Thomas Krause[2], Jörn Ostermann[2], and Jörg Bitzer[3]

[1] *Institut für technische und angewandte Physik GmbH, Marie-Curie-Str. 8, 26129 Oldenburg, Germany*
[2] *Leibniz Universität Hannover - Institut für Informationsverarbeitung, Appelstr. 9a, 30167 Hannover, Germany*
[3] *Jade Hochschule, Institute of Hearing Technology and Audiology (IHA), Ofener Str. 16-19, 26121 Oldenburg, Germany*

Correspondence should be addressed to Jörg Bitzer (`joerg.bitzer@jade-hs.de`)

ABSTRACT

Gunshot classification in audio files is used in forensics, surveillance, and multimedia analysis. In this contribution we show that it is possible to use data augmentation in order to enlarge the training set of a rare event like a gunshot with artificial data based on a simple but sufficient model, and a database of room impulse responses. The results indicate that the enlarged database increases the accuracy in a classification task significantly, even if no real data is used for training at all.

1 INTRODUCTION

Gunshot detection and classification is used in several application areas. In audio forensics the detection of gunshots is used to determine the timing of shot recordings [1] if multiple recordings exist, and in a second step to classify the weapon used [2]. Gunshot detection is also applied to analyze large multimedia databases, like TV shows or films to automatically detect scenes with gun violence [3]. Real time detection and classification is necessary, if gunshots and aggressive behaviour are used to steer cameras in live surveillance at public places [4, 5].

In order to detect gunshots several techniques were proposed. In [1] and [6] a template and cross-correlation (matched filter) approach was found to be a good choice. However, Ahmed et al. [7] reported that the false positive rate was high. This paper and several other publications use classification frameworks based on a plurality of audio features and classification algorithms (e.g. [3]). Mel frequency-cepstral coefficients (MFCC) are widely used as features, often combined with other spectral features [3]. Different filterbanks to summarize the energy features were proposed, including Gabor [8] and Wavelet transformations [9]. Tested classification algorithms include Gaussian Mixture Models (GMM) [10], Hidden Markov Models (HMM) [11], since they include variations over time, Support Vector Machines (SVM) [7] and deep learning algorihms (DNN) [12].

For all machine-learning based algorithms one major problem is the large amount of training data necessary, which is usually not available for rare events. Gun violence is not rare worldwide, but audio recordings of sufficient quality are. In this contribution we use typical features (e.g. Energy in Mel bands) and a standard SVM based classifier. However, we address the issue of insufficient data by modelling gunshots and additionally using several databases of impulse responses to create a huge amount of artificial training data in order to improve the detection rate.

2 Synthetic Shot Model

The basic time-domain signal of the shot sound according to Maher [13] is composed of three consecutive components (see Figure 1),

1. the shock-wave $s(k)$
2. a pause signal $b(k)$ and the
3. muzzle blast $m(k)$.

The shock wave

$$s(k) = \begin{cases} 0 & \text{for } k \in \{1, M_{SW}\} \\ A_{SW} - (k-2)\frac{2A_{SW}}{M_{SW}-3} & \text{for } 1 < k < M_{SW} \end{cases} \tag{1}$$

is modelled by an impulse of duration T_{SW} with a sampling rate f_s. A_{SW} denotes the linear amplitude of the signal, $k = 1...M_{SW}$ is the sample index, and $M_{SW} = f_s \cdot T_{SW}$ the number of samples used.
The shock wave is followed by the silent signal $b(k) = 0$ with duration $T_{SP} f_s$ and the amplitude $A_{SP} = 0$. This silence is followed by the muzzle blast with the duration T_{MB}. According to the Friedlander equation [14] in the discrete time domain this is defined by:

$$m(k) = A_{MB} \cdot \left[1 - \frac{k-1}{R_0} \exp^{-w(k-1)/R_0}\right] \tag{2}$$

where A_{MB} denotes the linear amplitude, $M_{MB} = f_s \cdot T_{MB}$ the number of samples, R_0 the length of the positive amplitude segment and w the rate of the exponential decay.

In the synthesis algorithm the length of the positive amplitude segment R_0 is determined by the parameter v_{MB} depending on the total length of the muzzle blast M_{MB}:

$$R_0 = v_{MB} \cdot M_{MB} \tag{3}$$

with $0 < v_{MB} \leq 1$. In the continuous time domain, it is defined as follows:

$$T_0 = v_{MB} \cdot T_{MB}. \tag{4}$$

Summarized, the full gunshot model is given by

$$g(k) = \begin{cases} s(k) & \text{for } 1 \leq k \leq M_{SW} \\ b(k - M_{SW}) & \text{for } M_{SW} < k \leq \\ & (M_{SW} + M_{SP}) \\ m(k - (M_{SW} + M_{SP})) & \text{for } (M_{SW} + M_{SP}) < k \leq \\ & (M_{SW} + M_{SP} + M_{MB}) \end{cases} \tag{5}$$

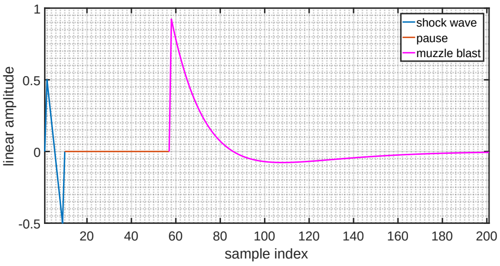

Fig. 1: Exemplary representation of a synthetic basic signal of a shot sound: shock wave (blue), pause (red) and muzzle blast (magenta)

Table 1: Variable parameters of the basic signal for the synthetic generation of shot sounds

Model segment	Parameter description	Parameter
shock wave $s(k)$	amplitude	A_{SW}
	duration in s	T_{SW}
silence $b(k)$	duration in s	T_{SP}
muzzle blast $m(k)$	amplitude	A_{MB}
	duration in s	T_{MB}
	proportion of positive amplitude segment	v_{MB}
	Rate of exponential decay	w

Figure 1 shows a synthetic basic signal of a shot sound without reverberation, which was created according to equation 5 with $A_{SW} = 0.5$, $T_{SW} = 0.003$ s, $T_{SP} = 0.001$ s, $A_{MB} = 1$, $T_{MB} = 0.003$, $v_{MB} = 0.2$, $w = 1.3$ and a sampling frequency of $f_s = 48$ kHz.

In Table 1 all individual model components with the associated variable parameters are given, which shows the possible parameter space. In order to simulate gunshots in a more natural surrounding, the anechoic model is convolved with room impulse responses $h_j(k)$:

$$d_{i,j}(k) = g_i(k) * h_j(k). \tag{6}$$

The convolution with a large number of different room impulse responses (RIR) increases the variance of the real-sounding synthetic shot sounds $d_i(k)$.
For the development and testing of the algorithm, room impulse responses out of three databases were used:

- *Aachen Impulse Response Database Version 1.4* (AIRD *1.4*)[15]

- *Multichannel Acoustic Reverberation Database at York* (MARDY)[16]
- *Open Air Impulse Respones Database* (OAIRD)[17]

All RIRs were adjusted to a fixed sampling rate by resampling ($f_s = 48$ kHz) and normalized to scale the final reverberant shot sound by the amplitude of the basic model.

3 Classification Framework

In this contribution we focus on the merits of an extended database for training. Therefore, a simple two-class problem will be used for classification: *Shot sound* or *no shot sound*. All training and test data are downsampled to a sampling frequency of 16 kHz. Data blocks of 512 samples (32 ms) are extracted from each test and training signal. For training or test data labeled as a gunshot the maximum of the gunshot is centered in the block. This in-block positioning is a-priori information that would not be available in a real-world experiment. For all no shot sounds, the signal block with length of 512 samples is determined randomly.

Every extracted data block is divided into three blocks with a length of 256 samples and an overlap of 50%. From each of the three blocks, two features are extracted, which are summarized chronologically in a feature vector. The first feature is the mel spectrum with ten sample points in the frequency range from 0 Hz to 8 kHz. As a second feature, the slope of the regression line of the signal's magnitude is extracted. With three audio blocks, this results in a vector of 33 features per training or test signal. Extraction of the second feature from an audio block is shown schematically in figure 2. The whole training feature matrix is linearly scaled between -1 and 1. The necessary scaling factors are used to scale the test data accordingly.

For the classification algorithm we used a support vector machine algorithm with a Gaussian Radial Basis Function (RBF) as the kernel from the library LIBSVM [18] .

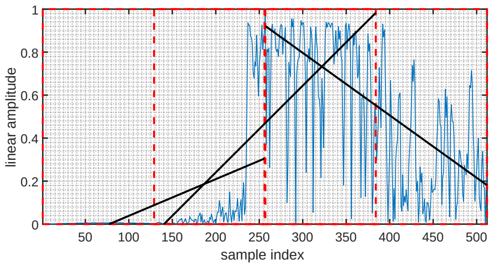

Fig. 2: Regression line (black) of three extracted overlapped audio blocks (red, dashed) of the magnitude of a normalized shot sound from the *UrbanSound8K* database (signal source: `https://serv.cusp.nyu.edu/projects/urbansounddataset/`, usage enabled by: `https://creativecommons.org/licenses/by/3.0/`)

Table 2: Number of audio data of the noise classes of the *UrbanSound8K* database

Noise class	Number
Shots	374
Air conditioners	1000
Car horns	429
Children playing	985
Dog barking	978
Drilling	1000
Engine idling	1000
Jackhammer	1000
Sirens	920
Street music	988
Total	8674

4 Test Methodology

For our classification task we used the *UrbanSound8K*[19] audio database. In addition to 374 real shot sound recordings, the database includes nine additional noise classes. The exact distribution and number of data is shown in Table 2.

4.1 Extension Of training data by synthetic gunshots

For the extension of the training data in the different classification scenarios, a synthetic dataset with 2187 synthetic shot sounds without reverberation was generated. For each of the seven parameters of the shot

Table 3: Parameter values for generating the synthetic shot sounds without reverberation

Model segment	Parameter	Values
shock wave	A_{SW}	$\{0.10,\ 0.55,\ 1.00\}$
	T_{SW} in s $\cdot 10^{-3}$	$\{0.10,\ 0.15,\ 0.20\}$
pause	T_{SP} in s $\cdot 10^{-3}$	$\{0.090,\ 4.045,\ 8.000\}$
muzzle blast	A_{MB}	$\{0.10,\ 0.55,\ 1.00\}$
	T_{MB} in s $\cdot 10^{-3}$	$\{3,\ 4,\ 5\}$
	v_{MB}	$\{0.1,\ 0.3,\ 0.5\}$
	w	$\{1.0,\ 1.5,\ 2.0\}$

sound model, three realistic parameter values were determined and combined according to the literature of Beck, Maher and Shaw [20][21][14]. The parameter values are shown in table 3.

4.2 Test scenarios

In the following four different classification scenarios are presented. In each scenario, the training set, consisting of initially 187 real shots and 8413 noises from other sound classes, were augmented by different sets of synthetic shot sounds.

Scenario 1: Extension of the training data by 2187 synthetic shot sounds without reverberation.
Scenario 2: Extension of the training data by 2187 synthetic shot sounds with one room simulation. This scenario was repeated with three different room simulations (in the following: *Scenario 2.1, 2.2, 2.3*).
Scenario 3: Extension of training data by 8748 synthetic shot sounds consisting of 2187 synthetic shot sounds without reverberation and 6561 synthetic shots by simulating the reverberation-free shot sounds in three surroundings.
Scenario 4: Extension of training data by 24057 synthetic shot sounds consisting of 2187 synthetic shot sounds without reverberation and 21870 synthetic shots by simulating the reverberation-free shot sounds in ten surroundings.

In order to show the effect of the artificial gunshot sounds, in the training set we reduced the number of real gunshot sounds gradually down to zero. Each experiment is labeled by the reduction step R. Table 4 shows the distribution of real shot sounds in the training and test sets, depending on the reduction steps.

Table 4: Number of audio data of the noise classes of the *UrbanSound8K* database

Sound class	R	Number in training set	Number in test set
no shot	1...14	8413	187
shot	1	187	187
	2	168	187
	3	150	187
	4	131	187
	5	112	187
	6	94	187
	7	75	187
	8	56	187
	9	37	187
	10	19	187
	11	9	187
	12	2	187
	13	1	187
	14	0	187

4.3 Classification task

In order to be independent of the a-priori probability of the two classes, we decided that each classification test task has 187 shot sounds and 187 other sounds randomly drawn from the *UrbanSound8K*. For each reduction step R we computed 100 test runs and averaged the results. All results are reported as the accuracy of the classification task. Therefore, the worst result is 50% accuracy, which is the chance level for equally distributed test sets in a two-class problem.

5 Results

Figure 3 shows the results for the classification tasks for all scenarios.

In *scenario 1* the results show that already with an extension of the training set with reverberation-free synthetic shot sounds the detection rate could be improved. In the event that there were no real shot sounds in the training set, an accuracy of 70.39% could be achieved and thus a performance increase of 20.39%. From the results of *scenarios 2.1* to *2.3*, we can see that the convolution with just one RIR further improves the results. However, the improvement depends on the used RIR. The best result achieves an accuracy of 90.1% in the event that there are no real shot sounds in

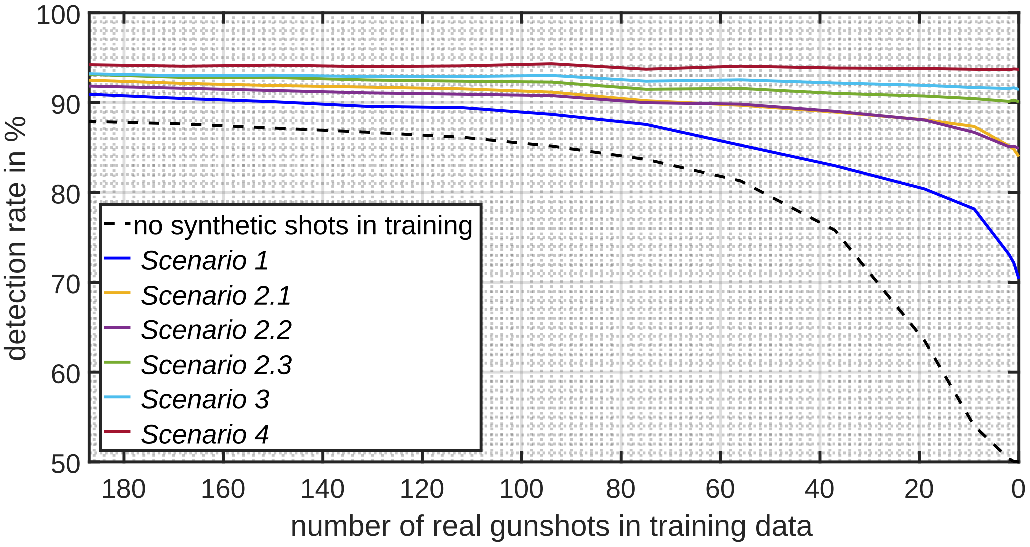

Fig. 3: Accuracy vs. number of real gunshot sounds for different scenarios. All results are averaged over 100 test runs

the training set.

Scenarios 3 and *4* show how the detection rate increases with the number of room simulations. In scenario 3 a detection rate of 91.47% could be achieved in the event that there are no real shot sounds in the training set. The results from *scenario 4* show that the detection rate is almost independent of the number of real shot sounds in the training set. The maximum detection rate of 94.22% for the event that there are 187 real shot sounds in the training set is only 0.5% above the detection rate calculated for the case without real shot sounds in the training set.

6 Final test: Comparing real and synthetic shots

In a final test we wanted to show, if real and synthetic shot sounds can be differentiated by the classification algorithm according to section 3. To show this, the classification algorithm was trained and tested with real and synthetic shot sounds, in which synthetic and real shot sounds were labeled differently. Three experiments were carried out with different synthetic datasets. In each experiment, 187 real and synthetic shot sounds were randomly drawn from the respective datasets and used as a test set. The rest of the data was used as training data. This process was repeated 100 times per experiment, and then the accuracy was averaged. As real shot sounds the 374 shot sounds from the *UrbanSound8K* database were used.

In *experiment 1*, 2187 synthetic shot sounds were used with one room simulation as a synthetic data set. In *experiment 2*, 8748 synthetic shot sounds were used as the synthetic data base. These are composed of 2187 anechoic synthetic shot sounds and 6561 synthetic shot sounds generated by simulation of the anechoic shot sounds in three different surroundings. In *experiment 3*, the synthetic data set from experiment 2 was extended by 15309 synthetic shot sounds through seven further room simulations. Table 5 shows the results. Especially with a huge amount of synthetic data in the training set, the accuracy is close to chance level which indicates that the differences are very small.

Table 5: Accuracy for the classification experiments 1, 2 and 3 averaged over 100 test runs

Experiment	Accuracy in %
1 (Synthetic shots)	78.84
2 (+ 3 RIRs)	74.48
3 (+ 10 RIRs)	57.53

7 Conclusions

In this contribution we showed that gunshot classification algorithms based on SVM can benefit from extending the training set with artificial data. Data augmentation is necessary, since the database for rare events are often small compared to the no-event databases. However, the model has to be good and the parameter space sampling should be diverse to get optimal results. It seems that the model established by Maher and the reported parameters are a good starting point for an artificial model. The most surprising result was that even with no real gunshot data at all in the training set a high accuracy of 90% in the *UrbanSound8k* database was achieved.

References

[1] R. C. Maher, E. Hoerr, "Audio Forensic Gunshot Analysis and Multilateration," presented at the *Audio Engineering Society Convention 145* (2018).

[2] E. Kiktova, M. Lojka, M. Pleva, J. Juhar, A. Cizmar, "Gun type recognition from gunshot audio recordings," presented at the *Biometrics and Forensics (IWBF), 2015 International Workshop on*, pp. 1–6 (2015).

[3] A. Pikrakis, T. Giannakopoulos, S. Theodoridis, "Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks," presented at the *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 21–24 (2008).

[4] C. Clavel, T. Ehrette, G. Richard, "Events detection for an audio-based surveillance system," presented at the *2005 IEEE International Conference on Multimedia and Expo*, pp. 1306–1309 (2005).

[5] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," presented at the *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 21–26 (2007 Sep.), doi:10.1109/AVSS.2007.4425280.

[6] A. Chacon-Rodriguez, P. Julian, L. Castro, P. Alvarado, N. Hernández, "Evaluation of gunshot detection algorithms," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 2, pp. 363–373 (2011).

[7] T. Ahmed, M. Uppal, A. Muhammad, "Improving efficiency and reliability of gunshot detection systems," presented at the *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 513–517 (2013).

[8] J. T. Geiger, K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," presented at the *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pp. 714–718 (2015).

[9] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *IEEE Transactions on information forensics and security*, vol. 3, no. 4, pp. 763–775 (2008).

[10] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, A. Sarti, "Scream and gunshot detection in noisy environments," presented at the *Signal Processing Conference, 2007 15th European*, pp. 1216–1220 (2007).

[11] I. L. Freire, J. A. Apolinario Jr, "Gunshot detection in noisy environments," presented at the *Proceeding of the 7th International Telecommunications Symposium, Manaus, Brazil*, pp. 1–4 (2010).

[12] H. Lim, J. Park, K. Lee, Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," Tech. rep., DCASE2017 Challenge, Tech. Rep (2017).

[13] R. C. Maher, "Modeling and signal processing of acoustic gunshot recordings," presented at the *Digital Signal Processing Workshop, 12th-Signal Processing Education Workshop, 4th*, pp. 257–261 (2006).

[14] S. Beck, H. Nakasone, K. Marr, "Variations in recorded acoustic gunshot waveforms generated by small firearms," *J. Acoust. Soc. Am.*, vol. 129, no. 4, pp. 1748–1759 (2011).

[15] I. of Communication Systems, "Aachen Impulse Response Database," URL `https://www.iks.rwth-aachen.de/`

en/research/tools-downloads/databases/aachen-impulse-response-database/, last acces: 29.01.2018.

[16] Speech, A. P. Laboratory, "MARDY (Multichannel Acoustic Reverberation Database at York) Database," URL https://www.commsp.ee.ic.ac.uk/~sap/resources/mardy-multichannel-acoustic-reverberation-database-at-york-database/, last acces: 29.01.2018.

[17] Openair, "Impulse response database," URL http://www.openairlib.net/auralizationdb, last acces: 29.01.2018.

[18] C.-C. Chang, C.-J. Lin, "LIBSVM – A Library for Support Vector Machines," URL https://www.csie.ntu.edu.tw/~cjlin/libsvm/, last acces: 29.01.2018.

[19] J. C. Salamon, J., J. P. Bello, "URBANSOUND8K DATABASE," URL https://urbansounddataset.weebly.com/urbansound8k.html, last acces: 29.01.2018.

[20] R. Maher, "Summary of Gun Shot Acoustics," *Montana State University* (2006).

[21] R. Maher, "Deciphering Gunshot Recordings," *Montana State University* (2008).