Maximum Likelihood Decision Fusion for Weapon Classification in Wireless Acoustic Sensor Networks

Héctor A. Sánchez-Hevia, *Student Member, IEEE*, David Ayllón, *Member, IEEE*, Roberto Gil-Pita, *Member, IEEE*, and Manuel Rosa-Zurera, *Senior Member, IEEE*

***Abstract*—Gunshot acoustic analysis is a field with many practical applications, but due to the multitude of factors involved in the generation of the acoustic signature of firearms, it is not a trivial task. The main problem arises with the strong spatial dependence shown by the recorded waveforms even when dealing with the same weapon. However, this can be lessen by using a spatially diverse receiver such as a wireless acoustic sensor network. In this work, we address multichannel acoustic weapon classification using spatial information and a novel decision fusion rule based on it. We propose a fusion rule based on maximum likelihood estimation that takes advantage of diverse classifier ensembles to improve upon classic decision fusion techniques. Classifier diversity comes from a spatial segmentation that is performed locally at each node. The same segmentation is also used to improve the accuracy of the local classification by means of a divide and conquer approach.**

***Index Terms*—Decision fusion, gunshots, multi-channel classification, wireless acoustic sensor networks.**

I. Introduction

WEAPON acoustic analysis has practical applications in many fields such as forensics, security, gun control or military tactics to name a few. Thus, the acoustic signature produced by explosive propelled weapons has been subject of study for some decades now [1]–[3]. In recent years, this field has become more relevant mainly due to the development of sniper detection and localization systems [4] aided by sensor fusion techniques and Wireless Acoustic Sensor Networks (WASN) [5]–[7]. Renewed interest in this topic has produced multiple approaches to gunshot detection over the last decade, whereas, acoustic weapon classification has not been widely studied yet with only a few available precedents [8], [9]. Regarding detection, most proposals use a simple acoustic event detector followed by a classification stage (gunshot vs other sounds) using pattern recognition techniques such as Gaussian Mixture Models (GMM) or Support Vector Machines (SVM) [10]–[12] in conjunction with classic acoustic analysis features. The standard approach to obtain spatial information relies on array processing [13]. However, in the context of a WASN, precise node localization and inter-node synchronization are basic requirements (for array processing) that are not easily met, and they still are under active research [14], [15].

The biggest problem in this field is the strong dependence on the shooter's location and orientation shown by the recorded waveforms, mostly because the acoustic disturbance created by firearms is highly directional [16] and its short time duration makes it behave like an impulse. This way, the perceived sounds are heavily influenced by the environment so that two recordings of the same gunshot taken at two distant locations can be completely dissimilar. On the other hand, one of the main contributions of WASN is the increase on spatial diversity provided by the large area that can be covered by the sensors [17]. From a classification standpoint, an acoustic signal recorded at a certain distance from the source is going to be convoluted and mixed with unwanted noise, affecting the accuracy of the classification system. Nonetheless, WASNs can provide multiple observations of an acoustic event, making it possible to obtain a more robust classification by fusing the available data [18]. Multi-observation classification has similarities with Bootstrap Aggregation (bagging) [19]. Classic bagged classification techniques such as AdaBoost or Random Forests take advantage of an ensemble of "weak" classifiers in order to "boost" the classification using decision fusion. In our case, since we have multiple observations of a single acoustic event, we can fuse the decision taken by an ensemble of nodes to achieve higher accuracy than that of the best node [20] in the same way as bagged classification. Multi-channel acoustic event detection and classification is rapidly gaining attention. Some recent examples on the literature (all of them using decision fusion) include [21], [22] and [23].

We propose a multi-channel weapon classification system that makes use of spatial information for a novel Maximum Likelihood-based decision fusion rule that accounts the relative location of nodes as a weight for their decisions. The system is intended for opportunistic WASNs, where each node has access to one microphone and computing capabilities, but cannot be seen as a wireless microphone array due to hardware limitations. Take as an example a dynamic wireless network formed

Manuscript received July 8, 2016; revised October 24, 2016 and December 22, 2016; accepted February 7, 2017. Date of current version May 23, 2017. This work was supported in part by the Spanish Ministry of Economy and Competitiveness under Project TEC2015-67387-C4-4-R, funds Spain/FEDER, and in part by the University of Alcalá under Projects CCG2015/EXP-056 and CCG2016/EXP-033. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Nobutaka Ono. *(Corresponding author: Héctor A. Sánchez-Hevia.)*

H. A. Sánchez-Hevia, R. Gil-Pita, and M. Rosa-Zurera are with the Department of Signal Theory and Communications, University of Alcala, Madrid 28801, Spain (e-mail: hectoradrian.sanchez@uah.es; roberto.gil@uah.es; manuel.rosa@uah.es).

D. Ayllón is with the Department of Signal Theory and Communications, University of Alcala, Madrid 28801, Spain, and also with the R&D Department, Fonetic, Madrid 28037, Spain (e-mail: david.ayllon@uah.es).

Digital Object Identifier 10.1109/TASLP.2017.2690579

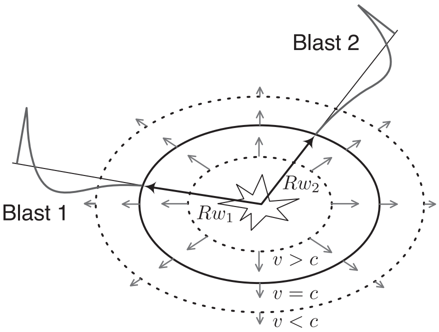


Fig. 1. Differences in blast wave radiation from an ellipsoid volume at two points.

by smartphones in which the lack of internode synchronization or the poor knowledge about the location of the nodes preclude the use of classic array processing. In this work, we are not considering the non-idealities of real wireless networks, however it is important to highlight that there are some problems regarding the transmission of node decisions, such as the presence of fading channels [24], faulty sensors [25] or transmission errors [26]. The proposed weighting method has some similarities with [27], where node weights are based on sensor-source distance.

Our local classification approach was first proposed on [28], it is based on a Divide and Conquer (D&C) strategy [29], the objective of which is to overcome the uncertainty produced by the lack of spatial references by taking advantage of simple problems to segment the space into a series of regions that are tackled independently. This spatial segmentation is later employed for the decision fusion scheme.

The paper is structured as follows: Section II briefly describes the acoustic model of gunshots; Section III presents the local classifier used by the nodes; on Section IV our decision fusion proposal is presented; finally, on Sections V and VI we describe the experiments, the obtained results and we extract some conclusions about the presented work.

II. Gunshot Acoustic Model

Explosive propelled weapons produce their characteristic sound as a result of the rapid expansion of gases at the end of their barrel, formally known as *muzzle blast*. *Weber's spectrum* is a model used to estimate the Fourier spectrum of blast waves on free air as a function of the radius of the expanding gas sphere [30], that is included in ISO norm 17201-2 [31]. According to this model, the energy of the explosion, thus, the radius of the gas sphere (Weber radius), is directly related to the wavelength of the blast. On firearms, the constraining effect of the barrel during the expansion affects the shape of the expelled gases making muzzle blasts highly directional. Directivity can then be explained as a relation between the listener location and the perceived Weber radius (R_w) [32]. Fig. 1 shows the differences in amplitude and time duration of the blast waves created by an ellipsoid volume at two points.

The second component is the *shock wave* created by supersonic projectiles. For a projectile with a speed $V > c$, defining *Mach number* as $M\mathrm{a} = V/c$, where c is the speed of sound,

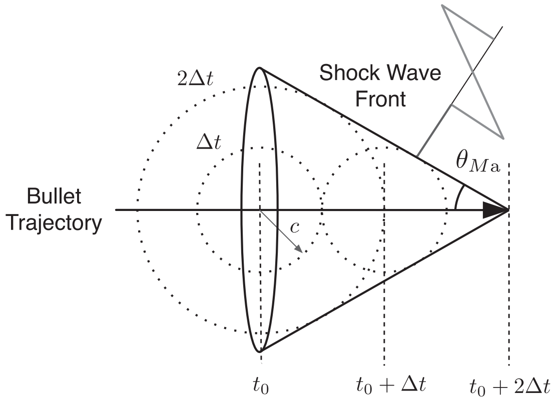


Fig. 2. Geometric model of shock wave propagation.

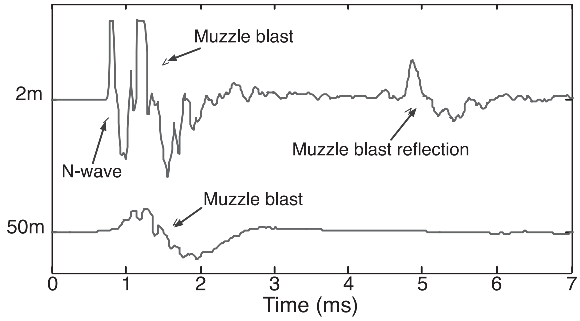


Fig. 3. Recorded waveforms of a .45 caliber handgun at two distant locations.

the generated shock wave propagates in conic shape forming an angle $\theta_{M\mathrm{a}} = \arcsin(1/M\mathrm{a})$ with the bullet trajectory as shown in Fig. 2. It is commonly called *N-wave* due to its characteristic geometry and, unlike the muzzle blast, it has a local influence (its energy is much lower) since it only appears at distances close enough to the trajectory of the bullet.

On a real scenario the recorded waveform may be very different from that described by the ideal model. In close range recordings, ground reflections from both muzzle blasts and shock waves, along with the sound produced by the firing mechanism of the weapon, are most likely overlapped with the direct signal. While, in long range recordings, the influence of the propagation path has a great impact on the received sound, due to its impulse-like components.

Additionally, non-idealities on the recording equipment can also produce some artifacts, the most relevant being signal saturation either at the microphone or at the analog front-end. Saturation is very likely to occur given the high sound pressure levels created by muzzle blasts that can exceed 140 dB.

Fig. 3 illustrates various of these effects with two recordings of a .45 caliber handgun. It is worth mentioning that the directivity of the muzzle blast combined with undesired acoustic phenomena, commonly makes the differences between recordings of the same weapon at distant locations greater than those of different weapons recorded at the same location.

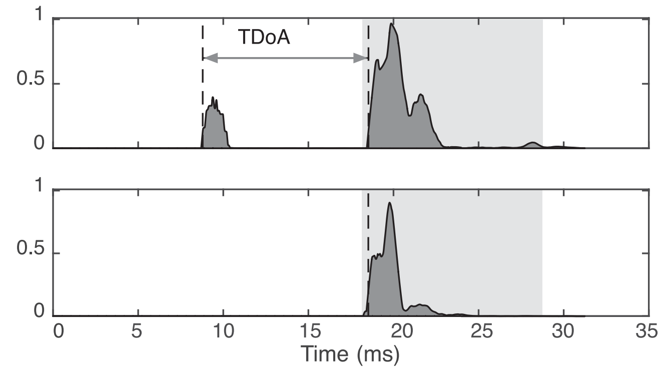


Fig. 4. Energy moving average of a gunshot recorded at two locations. Selected signal segment represented by a light gray area.

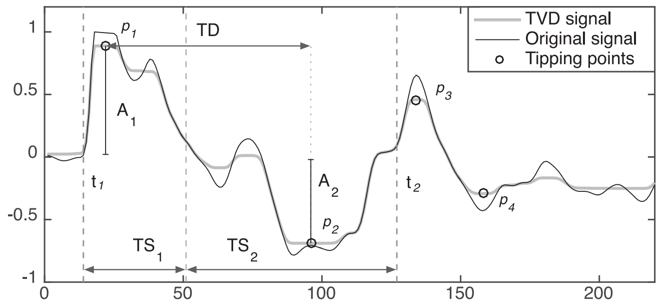


Fig. 5. Original signal, TVD processed signal and reference points for model-based feature extraction.

III. Acoustic Weapon Classification

This section describes the classification system used locally by the nodes. We are working with three broad categories: rifles, handguns and shotguns.

A. Feature Set

The proposed feature set is composed of 29 features: 22 standard features and 7 model-based features. It is computed using a signal segment $\mathbf{s}$ of length 10.7 ms (1024 samples at 96 kHz) that contains the Muzzle blast. This segment is selected from the starting point of the Muzzle blast which is detected using a moving average of the squared input signal with a rectangular window (length 0.67 ms) as shown in Fig. 4. The standard features are composed of two temporal features extracted from $\mathbf{s}$ and a series of spectral descriptors extracted from its *Fast Fourier Transform* (FFT) $\mathbf{z}$. From $\mathbf{s}$ we compute its energy level in decibels (as $20\log_{10}(\sum {s_n}^2)$) and its zero-crossing rate. From $\mathbf{z}$ we extract 4 spectral descriptors, namely: centroid, kurtosis, slope and roll-off [33]. $\mathbf{z}$ is also used to compute 16 Mel-frequency Cepstral Coefficients (MFCCs) [34].

Model based features are used to collect some temporal information that is not clearly reflected by standard features. Since the muzzle blast is always appearing (perfect detection is assumed) and it is the main source of energy, a secondary energy source preceding it with a lower energy level and shorter time duration must be an N-wave. (see Fig. 4 for a visualization of this statement). N-waves only appear with supersonic ammunition, and even then, the microphone has to be close enough to the trajectory of the projectile to capture it. From here, two conclusions can be made: weapons firing subsonic ammunition (such as shotguns) do not have N-waves and neither do those recordings taken from the back of shooter. On top of this, the longer the distance to the shooter the larger the Time Difference of Arrival (TDoA) between the N-wave and the muzzle blast. This information is obtained using the TDoA between energy clusters as a feature (see Fig. 4). In case only one cluster appears, it is set to a default value (zero).

The remaining model-based features are focused on the Muzzle blast waveform. They have more to do with the spatial aspect of the problem since the shape of the muzzle blast is a good range indicator (overlapping reflections at close range and convolution at long range). We propose to extract some shape descriptors using an algorithm that scans $\mathbf{s}$ in order to find its more prominent tipping points $\mathbf{p}$ (peaks and valleys). Fig. 5 illustrates this and the following concepts with an example taken from our database. The signal is first processed using Total Variation Denoising (TVD) [35]. TVD gets rid of the small variations in the signal but unlike low-pass filtering it preserves sharp edges. The positive and negative sections of the TVD processed signal are scanned independently using Matlab's *findpeaks* function (the negative section is first inverted). Local peaks are avoided by setting conditions to the peak finding algorithm (based on heuristics), specifically, minimum peak height and minimum distance between peaks. Once the tipping points have been found, taking the index value of the first two elements of $\mathbf{p}$ we use their Time Difference (TD) as a representative value of the duration of the first cycle and their amplitude ratio (A_1/A_2) as a symmetry measurement. We also find the zero-crossing points of the the first cycle and use them to compute its half cycle ratio (TS_1/TS_2) and its energy ($20\log_{10}(\sum_{n=t_1}^{t_2} {s_n}^2)$). Lastly, we include the number of tipping points and their maximum time difference as features, seeing that they are a good indicator of the presence of ground reflections.

A modified version of this feature set was first proposed in [28]. In that work the focus of the specialized features was set on the N-wave. We have since realized that muzzle blasts are a more reliable information source and have changed the approach accordingly, achieving a small decrease of the classification error rate.

B. Single-Channel Gunshot Spatial Information Extraction

In order to take advantage of spatial information without resorting to classic localization algorithms, we have reformulated some of the problems of the field, turning them into simpler problems that do not require the use of multiple information sources to be solved.

Using a single microphone, it is no longer possible to triangulate the shooter's location. However, some information can be obtained by making a rough estimation of his/her proximity to the node. In the current implementation, we are discerning between close range ($d < 20\,m$) and medium range ($d > 20\,m$) discharges, nevertheless, the proposed methodology is valid for

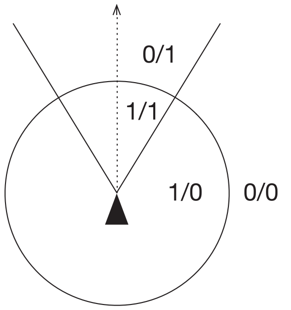


Fig. 6. Schematic representation of the spatial segmentation provided by the first classification stage (range/aligment).

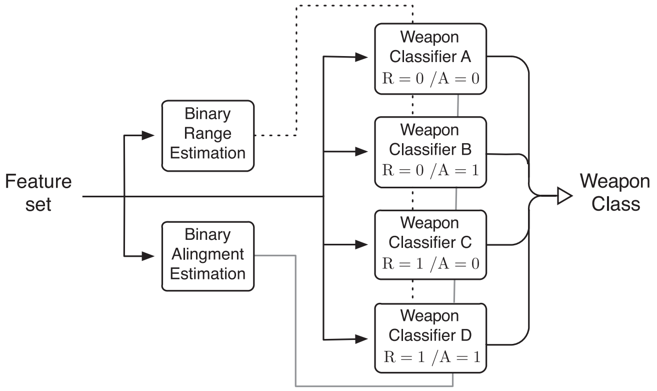


Fig. 7. Simplified diagram of the LS-LDA Classifier tree.

further subdivisions. Trajectory estimation also suffers from the lack of spatial references. It has been replaced by a rough estimate of the proximity of the sensor to the trajectory of the bullet into two broad alignment categories: on-axis and off-axis. On-axis implies that N-waves might appear at a given microphone location, while off-axis represents any other location.

This way each node has a certain degree of knowledge about its location in relation to the gunshot that can be exploited using a D&C strategy. This is performed by employing the outcomes of the range and alignment estimators to segment the space into four regions, according to Fig. 6.

C. LS-LDA Classification Tree

Since spatial influence on the signals is the main problem for classification, in [28] we first proposed to analyze different spatial regions independently. This is achieved with a classification tree that uses the outcome of two spatial estimators to segment the space by selecting between a set of localized weapon classifiers. The localized classifiers are designed using a specific subset of events, so that, they do not contemplate the existence of the other regions. In more correct terms we can say that the first stage is performing a division of the feature space (before weapon classification) based on a priori knowledge on the observation. A simplified diagram of the classifier tree is shown in Fig. 7. Notice that on the second stage only one localized classifier is active at a time, based on the values of R (range estimation) and A (alignment estimation).

We are using Least Squares Linear Discriminant Analysis (LS-LDA) [36] for every classifier in the tree instead of more "capable" solutions, because non-linear classifiers have shown an overfitting tendency when dealing with the presented problems, specially with weapon classification as we will see later. LS-LDA is a linear classification method, albeit it is possible to approximate nonlinear boundaries with a piece-wise linear function by assigning different classifiers to different regions of the feature space as the tree does.

The output of a LS-LDA for a given observation, is obtained as a linear combination of the feature vector $\mathbf{x}$ of length L, according to: $y = b + \sum_{n=1}^{L} w_n x_n$, where b is a bias value and w_n is the weight applied to the n-th feature. Multi-class LS-LDA is commonly tackled defining C binary classes and applying one-against-all, which entails that only one class is labeled as true for a given observation (one-hot encoding). This way, the output is a vector $\mathbf{y} = [y_1, ..., y_C]^T$ where each element represents the "score" of one of the classes. The final decision D is taken by selecting the largest element of $\mathbf{y}$ using:

$$D = \arg\max_c (y_c). \tag{1}$$

Let us consider a training set composed of N observations, where each element is $\mathbf{q}_i = [1, x_{i1}, ..., x_{i\mathrm{L}}]^T$, $i = 1, ..., N$. In matrix form, we can define an input matrix $\mathbf{Q}$ of dimensions $(L+1) \times N$ and a weight matrix $\mathbf{V}$ of dimensions $C \times (L+1)$, where C is the number of target classes, so that the output is computed as $\mathbf{Y} = \mathbf{V} \cdot \mathbf{Q}$. If we also define a binary target matrix $\mathbf{T}$ of dimensions $C \times N$ representing the desired outputs (true class labels in one-hot encoding), the error can be expressed as $\mathbf{E} = \mathbf{V} \cdot \mathbf{Q} - \mathbf{T}$. LS-LDA feature weights are computed in a single matrix operation with:

$$\mathbf{V} = \mathbf{T}\mathbf{Q}^T(\mathbf{Q}\mathbf{Q}^T)^{-1}. \tag{2}$$

In order to obtain the weight matrices for the classifier tree, let us begin with the first stage by defining two target vectors $\mathbf{t}_A$ and $\mathbf{t}_B$ of length N that represent the true range and alignment labels respectively. From these two vectors and input matrix $\mathbf{Q}$ the weights for the first stage are computed with:

$$\mathbf{v}_A = \mathbf{t}_A\mathbf{Q}^T(\mathbf{Q}\mathbf{Q}^T)^{-1}, \text{ and } \mathbf{v}_B = \mathbf{t}_B\mathbf{Q}^T(\mathbf{Q}\mathbf{Q}^T)^{-1}. \tag{3}$$

From where the range and alignment estimations of each observation are obtained as:

$$\mathrm{R}_i = \begin{cases} 1 \text{ if } \mathbf{v}_A\mathbf{q}_i > 0.5 \\ 0 \text{ if } \mathbf{v}_A\mathbf{q}_i \leq 0.5 \end{cases}, \text{ and } \mathrm{A}_i = \begin{cases} 1 \text{ if } \mathbf{v}_B\mathbf{q}_i > 0.5 \\ 0 \text{ if } \mathbf{v}_B\mathbf{q}_i \leq 0.5 \end{cases}. \tag{4}$$

Now, using the outputs of the first stage, let us make a partition of $\mathbf{Q}$ into four subsets $\mathbf{Q}_k$, $k = 1, ..., 4$ containing the observations of each of the four spatial regions and defined as:

$$\mathbf{Q}_1 = \{\mathbf{q}_i | \, \mathrm{R}_i = 0 \text{ and } \mathrm{A}_i = 0\}, \tag{5}$$

$$\mathbf{Q}_2 = \{\mathbf{q}_i | \, \mathrm{R}_i = 0 \text{ and } \mathrm{A}_i = 1\}, \tag{6}$$

$$\mathbf{Q}_3 = \{\mathbf{q}_i | \, \mathrm{R}_i = 1 \text{ and } \mathrm{A}_i = 0\}, \tag{7}$$

$$\mathbf{Q}_4 = \{\mathbf{q}_i | \, \mathrm{R}_i = 1 \text{ and } \mathrm{A}_i = 1\}. \tag{8}$$

The reason why we are using the outputs of the first stage instead of the true range and alignment labels for the division of the training set, is to lessen the restrictions on the segmentation by letting the classifiers select which observations belong where. This method of performing the segmentation is the only difference between the current implementation of the LS-LDA tree and the one introduced in [28].

Since each of the observations belongs to one of the 4 subsets and has a true weapon label associated, we can define matrix $\mathbf{T}_k$ as the target matrix of those observations within $\mathbf{Q}_k$, so that the weight matrices of the localized weapon classifiers are obtained with:

$$\mathbf{V}_k = \mathbf{T}_k {\mathbf{Q}_k}^T (\mathbf{Q}_k {\mathbf{Q}_k}^T)^{-1}, k = 1, ..., 4. \tag{9}$$

IV. Classification in WASN

From a data-fusion point of view, The spatial diversity provided by a WASN can be exploited either fusing the signals in the time domain with beamforming [37]; fusing them in the feature domain [38]; or fusing the individual decisions taken by the nodes. Our framework does not contemplate the use of WASN as a wireless microphone array due to hardware limitations, and so, beamforming is not feasible. In non-uniform ad-hoc WASNs with different devices from different manufacturers, or without a good communication layer management, synchronization of the sampling rates may be hard (or even impossible), and the resulting signal drift must then be taken into account by the signal processing algorithms [17]. Given that we are working on a distributed scenario, decision fusion is preferred since it minimizes data transmissions and has great scalability capabilities. Regardless of the number of nodes in the network the main processing remains the same; each node tries to classify the acoustic event and shares the results with the network. The only aspect notably affected by an ensemble size increase is the decision fusion operation, which usually is a less demanding computation than those carried out locally by the nodes.

A. Decision Fusion

As we have previously mentioned, on a spatially diverse scenario, the signal received at different locations is subjected to variations that cannot be predicted during the training stage, which entails a performance decrease. However, since there are multiple available observations of a single acoustic event, we can fuse the decisions taken by the nodes to increase accuracy. Fig. 8 shows a schematic representation of a hypothetical scenario with a gun being fired and the locations of some nodes forming a WASN.

Let us have an ensemble of M nodes (classifiers) providing an output $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M]$ with $\mathbf{y}_l = \mathcal{G}(\mathbf{x}_l)$, $l = 1, ..., M$, where $\mathcal{G}$ is a function shared by every node in the ensemble and $\mathbf{x}_l$ is a local observation of $\mathbf{x}$. The final decision is: $D = \mathcal{F}(\mathbf{y}_1, ..., \mathbf{y}_M)$, where $\mathcal{F}$ is the fusion method. Previous research has shown that the major factor for a better accuracy is the diversity in the classification and so, the fusion method is of secondary interest [39]. However, choosing an appropriate fusion method can further improve the performance of the ensemble.

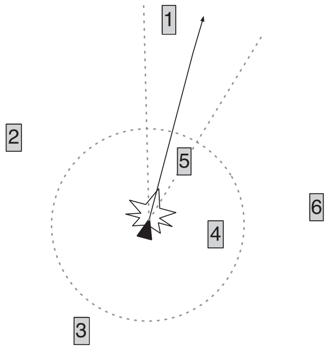


Fig. 8. Schematic representation of a gun being fired inside a WASN (nodes indicated by numbers).

Two of the most common fusion rules are average fusion and majority vote [40], although it goes without saying that there are more alternatives [41]. Average fusion is self explanatory; the outputs of the ensemble are averaged and treated as that of a single classifier. Majority vote works with local decisions, selecting the class label most represented among the M classifiers. These fusion rules are equivalent when the output of the classifier ensemble is already a local decision vector (there is no access to probabilistic outputs). If we were to perform one-against-all multi-class classification using average fusion, where the output of each classifier in the ensemble is a vector $\mathbf{y}_m = [y_{m1}, ..., y_{m\mathrm{C}}]^T$, with C being the total number of classes, the final decision D would be obtained with the following expression:

$$D = \arg\max_c \left(\frac{1}{M} \sum_{m=1}^{M} y_{mc} \right). \tag{10}$$

Since it is very likely for some of the nodes to perform better than others [42], it is only logical to assume that we should assess the contribution of each classifier to the ensemble according to some performance metric by adding a weight vector $\mathbf{w} = [w_1, ..., w_M]$. Doing this, the final decision becomes: $D = \mathcal{F}(w_1 \mathbf{y}_1, ..., w_M \mathbf{y}_M)$. There are many valid methodologies to compute $\mathbf{w}$, ranging from statistical analysis to heuristics. One common solution is to use the accuracy of each classifier to weight its contribution using:

$$D = \arg\max_c \left(\sum_{m=1}^{M} w_m y_{mc} \right),$$
$$\text{with: } w_m = 0.5 \log((1 - e_m)/e_m, \tag{11}$$

where e_m is the error rate of the m-th node.

In this work we propose a custom decision fusion weighting method based on Maximum Likelihood estimation that is explained below.

B. Maximum Likelihood Decision Fusion

Let us consider that a set of estimations of the likelihood of a given event belonging to each class is available, where the estimation of each node has an associated error that depends on its relative location to the source, which classifier was used and the true class of the event. The objective is to obtain the best possible estimation of the true class of the event by fusing the available information. Let us also consider that the outputs of the m-th node $\mathbf{y}_m = [y_{m1}, ..., y_{m\mathrm{C}}]$ follow a gaussian distribution so that its Probability Density Function (PDF) is given by:

$$f(\mathbf{y}_m;\mathbf{t},\mathbf{C}_m) = \frac{1}{\sqrt{(2\pi)^C|\mathbf{C}_m|}} e^{\left(-\frac{1}{2}(\mathbf{y}_m-\mathbf{t})^T\mathbf{C}_m^{-1}(\mathbf{y}_m-\mathbf{t})\right)}, \tag{12}$$

where $\mathbf{t}$ is the true class or target vector and $\mathbf{C}_m$ is the covariance matrix of the PDF. In this general case we consider different covariance matrices for different estimations.

At this point, it is possible to obtain the most likely decision using a Maximum Likelihood estimator of the true class of an event $(\mathbf{t})$ given a set of samples $\mathbf{y}_m$. Thus, the log-likelihood LL of a given estimation set is obtained with:

$$LL = \sum_{m=1}^{M} \log(f(\mathbf{y}_m;\mathbf{t},\mathbf{C}_m)). \tag{13}$$

Replacing equation (12) in (13) and simplifying, the following expression is obtained:

$$LL = \frac{1}{2}\sum_{m=1}^{M}\left(-\log((2\pi)^C|\mathbf{C}_m|) - (\mathbf{y}_m-\mathbf{t})^T\mathbf{C}_m^{-1}(\mathbf{y}_m-\mathbf{t})\right). \tag{14}$$

In order to maximize each component of the log-likelihood function, expression (14) has to be differentiated with respect to $\mathbf{t}$ and equaled to zero, leading to the following system of equations:

$$\sum_{m=1}^{M}\mathbf{C}_m^{-1}\mathbf{t} = \sum_{m=1}^{M}\mathbf{C}_m^{-1}\mathbf{y}_m \tag{15}$$

Now, $\mathbf{t}$ can be cleared from (15), arriving to :

$$\hat{\mathbf{t}} = \sum_{m=1}^{M}\left(\sum_{m=1}^{M}\mathbf{C}_m^{-1}\right)^{-1}\mathbf{C}_m^{-1}\mathbf{y}_m, \tag{16}$$

where the final decision is:

$$D = \arg\max_c(\hat{t}_c), \tag{17}$$

Notice that the hat indicates that $\hat{\mathbf{t}}$ is an estimate. This is an implication of the assumptions taken to arrive at this conclusion that are not necessarily strictly true.

Finally, we can extrapolate this conclusion to define a fusion rule for our classifier ensemble. Node weights are obtained with:

$$\mathbf{W}_m = \left(\sum_{m=1}^{M}\mathbf{C}_m^{-1}\right)^{-1}\mathbf{C}_m^{-1}, \tag{18}$$

where $\mathbf{W}_m$ is a $C \times C$ matrix and the weights for each node and class are found using all the covariance matrices of the ensemble. Defining an auxiliary vector $\mathbf{z}_m = \mathbf{W}_m\mathbf{y}_m$, the final decision is:

$$D = \arg\max_c\left(\sum_{m=1}^{M} z_{mc}\right). \tag{19}$$

In the case of a multi-observation classification ensemble, where all the observations are evaluated with a single classifier and the influence of node locations has not been characterized, we have $\mathbf{C}_m = \mathbf{C}$, thus, the presented fusion method becomes equivalent to average fusion. Defining $\mathbf{I}$ as the identy matrix and $\mathbf{B} = \mathbf{C}^{-1}$, we know that $\mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$ and $(\alpha\mathbf{B})^{-1}\mathbf{B} = (1/\alpha)\mathbf{I}$, making $w_m = 1/M$. This result does not hold true when there is some degree of known error diversity and $\mathbf{C}_m$ takes different values.

C. *Proposed Location-Driven Decision Fusion*

In order to take advantage of the ML decision fusion, we first need to characterize the covariance matrix of each node on the network. Let us assume a hypothetical scenario where both the location of the nodes and the sources is fixed. It would then be possible for each node to use its own classifier trained for that particular scenario, from where $\mathbf{C}_m$ could be found. On a real scenario however, the relative position of the sources and the nodes is prone to change affecting the classification accuracy. Albeit, if we were able to characterize the relation between the classifier output and the relative location of the node we could still estimate $\mathbf{C}_m$ as a function of the location, even when using a single classifier for the whole ensemble similarly to [27]. For the time being, we are performing a discrete spatial division as described in Section III-B where the spatial relation is estimated by each node between 4 possibilities. For each region, the LS-LDA tree has an associated covariance matrix $\mathbf{C}_k$, $k = 1, .., 4$. This entails that, for every new gunshot, each node assigns $\mathbf{C}_m$ to one of the possible $\mathbf{C}_k$ based on this selection. Notice that this solution is scalable, the more spatial information available about the source, the greater the spatial segmentation can be.

In order to compute $\mathbf{C}_k$ for the present system, lets consider an output matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M]$ and a target matrix $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_M]$ so that the estimation error is $\mathbf{E} = \mathbf{Y} - \mathbf{T}$. The sample covariance of the estimation is then:

$$\mathbf{C} = \frac{1}{N}(\mathbf{Y}-\mathbf{T})(\mathbf{Y}-\mathbf{T})^T = \frac{1}{N}\mathbf{E}\mathbf{E}^T, \tag{20}$$

replacing in this expression with the variables from III-C we can estimate the covariance matrix of the specialized classifiers as $\mathbf{C}_k = \frac{1}{N}\mathbf{E}_k{\mathbf{E}_k}^T$, where $\mathbf{E}_k = \mathbf{V}_k\mathbf{Q}_k - \mathbf{T}_k$.

Since training the classifier using a mean-square-error criterion gives outputs that approximate posterior class probabilities [43] it is good practice to saturate the outputs using the target interval as limits (probability 0 to 1) before computing the error.

We have found that for those scenarios where computational complexity is an issue, it is possible to discard the values of $\mathbf{C}_k$ that are not part of the main diagonal which is equivalent to assuming uncorrelated classes. While this method yields slightly worse results, it has the benefit of turning matrix operations into element wise operations, since by doing this, matrix $\mathbf{W}_m$ becomes vector $\mathbf{w}_m$. This is specially relevant for the inversion of the sum. Given that the values of $\mathbf{C}_k$ are fixed during training, they can be stored as $\mathbf{B}_k = \mathbf{C}_k^{-1}$, making the inversion of the sum the only matrix inversion computed during run time. Expanding further into computational complexity, it is possible to completely avoid matrix inversions by storing all the possible values of $(\sum_{m=1}^{M} \mathbf{C}_m^{-1})^{-1}$ on a lookup table. In a system with M maximum nodes and K possible covariance matrices, adding a zero matrix $\mathbf{C}_0$ as an additional covariance matrix, the table length can be found as the number of combinations with repetition using:

$$\binom{(K+1)+M-1}{M} - 1 = \frac{(K+M)!}{M!K!} - 1, \qquad (21)$$

where $\binom{a}{b}$ represents the combinatorial number. The table is valid for any WASN with a number of nodes less or equal to M, since $\mathbf{C}_0$ is added in order to consider null contributions to the sum. This solution is a clear tradeoff between memory and computing power, so whether or not it is advisable to use a lookup table will depend on the target platform. With $M = 8$ and $K = 4$ there are 494 possible combinations. If $C = 3$ and considering single-precision (32-bit) floating-point representation, the table would require 17.37 kilobytes of memory.

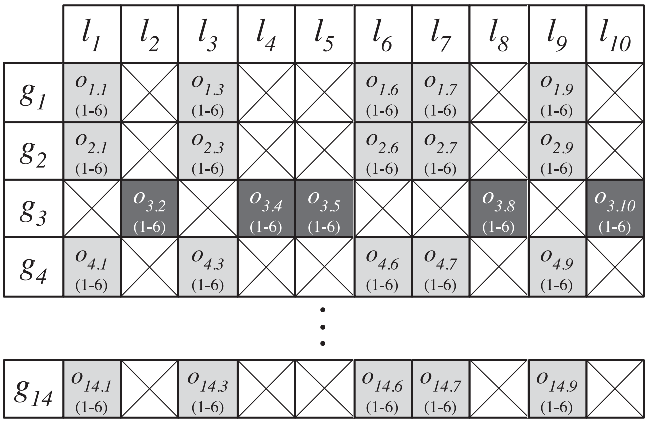
V. Experiments and Results

After introducing the specifics of the presented classification system, in this section we will describe the experimental setup and the obtained results.

A. Database Description

The sounds that make up the database are commercially available as part of a sound library offered by the company *BOOM Library* under the name *"GUNS - Construction Kit"* [44]. All the signals are recorded at 96000 Hz using various high-quality microphones and recording equipment. Not every sound available on the library was included on the database, some sounds were discarded due to inconsistency issues and for the sake of obtaining a balanced database. While the sound library description provides some valuable information about the microphone locations it is not detailed enough to tackle array processing, on top of this, various microphone types were used in the recordings having differences between them in terms of frequency response and directivity. While having different microphones can be beneficial towards diversity, it is far from ideal for array processing, where the common approach is to assume identical microphones.

The database contains recordings of 14 weapons: 5 handguns, 5 rifles and 4 shotguns. There are 6 recordings (shot repetitions) of every weapon at 10 different locations (6 observations per weapon-location combination), adding up to a total of 840 gunshots. Of the 10 unique locations, 4 are labeled as short-range and 6 as medium range, whereas 6 are labeled as on-axis and 4 as off-axis. N-waves only appear in 22.1% of the recordings, not appearing at all for 6 of the weapons (2 handguns and every shotgun) since they use subsonic ammunition.



	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	l_{10}
g_1	$o_{1.1}$ (1-6)	☒	$o_{1.3}$ (1-6)	☒	☒	$o_{1.6}$ (1-6)	$o_{1.7}$ (1-6)	☒	$o_{1.9}$ (1-6)	☒
g_2	$o_{2.1}$ (1-6)	☒	$o_{2.3}$ (1-6)	☒	☒	$o_{2.6}$ (1-6)	$o_{2.7}$ (1-6)	☒	$o_{2.9}$ (1-6)	☒
g_3	☒	$o_{3.2}$ (1-6)	☒	$o_{3.4}$ (1-6)	$o_{3.5}$ (1-6)	☒	☒	$o_{3.8}$ (1-6)	☒	$o_{3.10}$ (1-6)
g_4	$o_{4.1}$ (1-6)	☒	$o_{4.3}$ (1-6)	☒	☒	$o_{4.6}$ (1-6)	$o_{4.7}$ (1-6)	☒	$o_{4.9}$ (1-6)	☒
⋮										
g_{14}	$o_{14.1}$ (1-6)	☒	$o_{14.3}$ (1-6)	☒	☒	$o_{14.6}$ (1-6)	$o_{14.7}$ (1-6)	☒	$o_{14.9}$ (1-6)	☒

Fig. 9. Representation of the r-th division of the database with $n = 3$. Light gray: Train set, dark gray: Test set.

B. Experiment Description

Our objective is to test the system in as close to real conditions as possible. This implies that, for any given gunshot, both the shooter's location and the fired weapon are going to be new to the system. Since our database is not large enough for the complexity of the problem, a simple division of the data into a train set and test set is not efficient. In order to obtain the presented results we used a hybrid cross validation method that mixes Leave-One-Out Cross-Validation (LOOCV) [45] and *Repeated random sub-sampling validation*. In LOOCV a single observation is used as the test set and the remaining observations $(N - 1)$ as the training set. The process is repeated N times, until every observation has been tested, and then, the results are averaged to obtain the final performance metrics. Repeated random sub-sampling validation involves r random divisions of the data set into two equally sized and disjoint train and test sets with the consequent result averaging. We are also using random replacement to keep the train set balanced. Every iteration, b random sounds of the train set that belong to the tested class are duplicated, b being the number of observations in the test set.

In every iteration the database is randomly divided into a train set and a test set equally sized in terms of included positions (5 positions per set). Weapons are tested using LOOCV so that the test set is formed by the sounds of one gun at 5 positions while the training set contains the sounds of the remaining weapons at the remaining positions (13 guns × 5 positions). This is done in order to maximize the information available to the classifier about weapon classes while testing it with a previously "unheard" gun recorded at unknown locations. For each of the guns, r random location-wise database divisions are tested. It is important to remark that since there are only 5 microphone locations per division we have decided to count the 6 observations available per gun-location ($\mathbf{o}_{g,l}$) as additional microphone locations. Treating repetitions as different locations is justified by assuming they are recordings of the same gunshot taken at 6 locations very close to each other. In total we consider 30 possible locations (5 locations × 6 repetitions).

The final step involves the generation of p random permutations of the test set without repetition, that is, sorting the

TABLE I
CLASSIFICATION ERROR FOR DIFFERENT ENSEMBLES OF M CLASSIFIERS USING AVERAGE FUSION RULE

	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 6$	$M = 7$	$M = 8$
LS-LDA tree	30.2%	22.3%	16.9%	13.9%	**11.7**%	**10.2**%	**8.9**%	**7.9**%
LS-LDA	35.8%	26.9%	22.1%	19.3%	17.2%	15.7%	14.3%	13.5%
MLP	32.9%	23.6%	19.8%	16.8%	14.9%	13.3%	12.2%	11.3%
SVM lin	31.4%	23.7%	19.2%	16.9%	14.8%	13.3%	11.9%	10.9%
SVM RBF	**29.4**%	21.5%	**16.2**%	13.9%	12.0%	10.5%	9.5%	8.6%
Random Forest	31.1%	**21.2**%	16.4%	**13.8**%	12.1%	10.9%	10.2%	9.6%
K-NN	36.8%	29.6%	24.1%	21%	19.1%	17.3%	14.9%	13.8%

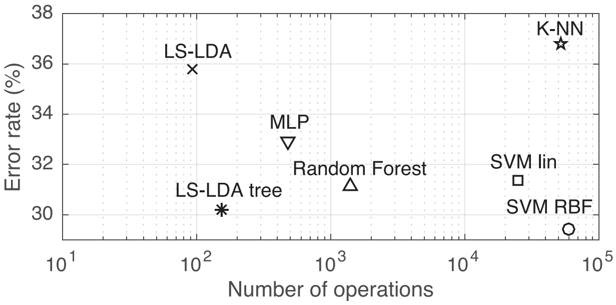


Fig. 10. Error rate (single channel) against number of operations (evaluation) for different classifiers.

observations included of the test set randomly. For each permutation we obtain the m-th ensemble error rate by fusing the decisions obtained for the first $m = 1, ..., M$ observations. The final results are computed averaging the errors obtained in the whole $14 \times r \times p$ experiments for each ensemble size. The presented data uses $r = 32$ and $p = 8$ adding up to a total of 3584 iterations.

Fig. 9 shows the r-th division for the third gun ($n = 3$), the sounds highlighted in light gray are used to train the classifiers while the sounds highlighted in dark gray are arranged in p random permutations in order to test the system. Each cell in the table contains the six observations available per gun-location combination. Notice that on each iteration half of the sounds remain unused.

The covariance matrices of the LS-LDA tree were obtained from the train set using LOOCV with weapon-location combinations (13 guns × 5 positions repetitions), leaving out all the available sounds of a gun at a single position and training with the rest. We opted to compute the train error using cross validation since the results are less prone to overfitting and so, it yields a closer estimation to the covariance values found when testing the system with new observations. As we have previously mentioned, the test set was never used in the design phase, including the estimation of the covariance matrices.

The features are normalized (between 0 and 1 according to the training set) prior to the training. For the experiments, first we compared the performance of the LS-LDA tree and some well established classifiers using the average fusion rule under equal conditions. Later, we compared the results obtained by the LS-LDA tree using different fusion rules. During the experiments, the exact same data set divisions were used for every classifier and fusion rule tested.

C. Discussion of the Results

Table I shows the results obtained for various classifiers using the average fusion rule. The classifiers are:

1) LS-LDA tree: The classifier described in Section III-C.
2) LS-LDA: single stage LS-LDA. Trained using Matlab's *fitcdiscr* function.
3) MLP: Multi-Layer Perceptron with 2 hidden layers (12 and 6 neurons) and 3 output neurons. Trained using Matlab's *patternnet* function.
4) SVM lin: SVM with a linear kernel. Trained in Matlab using *libSVM* [46] in ν-SVC mode with probability estimates enabled. Optimal value of hyperparameter ν is found using cross-validation in the same way as the LS-LDA tree covariance matrices.
5) SVM RBF: Same as above but using a Radial Basis Function (RBF) kernel. In addittion to ν, the optimal value of hyperparameter γ is also found using cross-validation.
6) Random Forest: A classifier ensembe formed by 128 decision trees. Trained using Matlab's *TreeBagger* function.
7) K-NN: K-nearest neighbors using L1 norm and the 5 closest neighbors. Trained with Matlab's *fitcknn* function.

We are using probabilistic outputs with every classifier and unless otherwise stated, the employed functions and libraries are set to their default configuration.

The results show a similar error decrease for every classifier when more nodes are added to the ensemble. This clearly points at spatial diversity as a major factor for achieving good accuracies, more relevant than the classifier itself. Nevertheless, among the tested classifiers, the LS-LDA tree obtained the best results for large ensembles while having a lower computational complexity than some of the other tested methods. In order to compare the computational cost of the tested classifiers we have calculated a naive approximation to the number of operations required to evaluate one observation with each classifier. We assume that memory operations are costless, that is, data is immediately accesible. In addition to this, we are giving the same cost to every operator: simple arithmetic operators (e.g. addition), relational operators (e.g. greater than), complex functions (e.g. exponentiation, square root, ...) and multiply-accumulate (MAC); they all take one operation. This is an oversimplification, although, it can be seen as an approximate to an implementation using speed optimizations, such as tabulated functions; and specialized hardware, such as a MAC unit. Fig. 10 shows the classification error (single node) plotted against the

TABLE II
CLASSIFICATION ERROR FOR THE LS-LDA TREE USING DIFFERENT FUSION RULES. STANDARD DEVIATION IN PARENTHESES

	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 6$	$M = 7$	$M = 8$
Maximum Likelihood	**30.2%** (4.0%)	**20.6%** (3.3%)	**15.0%** (2.8%)	**11.6%** (2.5%)	**9.5%** (2.5%)	**8.1%** (2.3%)	**7.0%** (2.2%)	**5.9%** (2.2%)
Average	30.2% (4.0%)	22.3% (3.4%)	16.9% (2.8%)	13.9% (2.7%)	11.7% (2.7%)	10.2% (2.6%)	8.9% (2.6%)	7.9% (2.5%)
Weighted Average	30.2% (4.0%)	22.7% (3.6%)	17.5% (3.2%)	14.5% (3.1%)	12.4% (3.2%)	11.0% (3.0%)	9.7% (2.9%)	8.7% (2.9%)
Majority Vote	30.2% (4.0%)	22.4% (3.1%)	17.2% (2.4%)	14.6% (2.8%)	12.8% (2.1%)	11.6% (2.7%)	10.6% (2.5%)	9.8% (2.5%)
Maximum Vote	30.2% (4.0%)	24.8% (4.2%)	19.8% (3.3%)	17.4% (3.7%)	14.5% (3.2%)	13.1% (3.4%)	11.7% (3.1%)	10.6% (3.1%)
Supervised Average	31.3% (3.3%)	22.0% (2.7%)	16.1% (2.4%)	13.2% (2.4%)	10.8% (2.4%)	9.5% (2.3%)	8.2% (2.3%)	7.2% (2.4%)

number of operations required for different classifiers (logarithmic scale).

Table II shows the results obtained using the LS-LDA tree when different decision fusion rules are applied:

1) Maximum Likelihood: fusion rule proposed in (19).
2) Average: Average fusion described in (10).
3) Weighted Average: weighted variation described in (11), with e_m computed from the train set using LOOCV in the same way as $\mathbf{C}_k$.
4) Majority Vote: D is assigned to the class label selected by the most nodes. Ties resolved by absolute maximum output.
5) Maximum Vote: hybrid between average and majority vote. Nodes vote for the winning class using their probabilistic output and assign a zero to the other classes.
6) Supervised Average: the weights for each zone and class are obtained using LS-LDA during training. In this configuration $\mathbf{w}_k$ is a 3 element vector out of 4 possible ones. The weight vector of each zone is computed with $M = 8$, for every training iteration the outputs of the nodes that have selected one particular zone are averaged, in case none of the M nodes did, the observation is withdrawn from the training set.

Of all the tested methods, only the supervised average and the proposed rule were able to improve upon average fusion. The proposed Maximum Likelihood fusion rule yields the best results, obtaining a similar error rate when 6 nodes are used to that of average fusion with 8 nodes. It is important to highlight that naive weighted fusion can be counterproductive, as shown in the table for one common weighted solution. Communication cost is fairly low regardless of the fusion method employed. Assuming a network where the outputs are transmitted in single-precision (32-bit) floating-point format, the total amount of data required to transmit the outputs of 8 nodes is just 768 bits per evaluation (without considering protocol overhead). With a weighted scheme, since every node has to have every weight stored, the only transmission needed is an index (per node) to represent the selected weight. Encoding said index as a single byte, the total transmission would increase by 64 bits. Since the system works with framed audio, the required bit rate with the present configuration is 78000 bps, although there is no need to take a decision every frame. In terms of computational complexity, the proposed fusion rule is by far the most demanding, with a number of operations proportional to M^2 while for every other method it is proportional to M. However Maximum Likelihood fusion can also be made proportional to M by using a tabulated implementation as suggested in Section IV-C.

The error rates obtained with the LS-LDA tree for range and alignment estimations are 3.5% and 5.8% respectively, which is in clear contrast with those obtained for weapon classification and it is thanks to this degree of accuracy that the spatial division is able to boost the system.

The obtained results show a strong relationship between the spatial resolution of the classifiers and the obtained error. From them, it is clear that the addition of spatial diversity to the system is of the utmost importance. With the LS-LDA tree, just by adding a second node in conjunction with the proposed fusion method, the classification error falls by almost 10% getting as low as 5.9% when 8 nodes are used. Keep in mind that the classifiers were trained without using any observations of the tested guns neither the tested locations, so it is safe to say that we are working in the most restrictive conditions that the current database allows.

In order to study the statistical significance level of the results, a paired-sample *t*-test has been carried out comparing the error probabilities obtained for the LS-LDA tree using the proposed Maximum Likelihood fusion rule with those obtained using the average fusion rule. Differences are statistically significative ($p < 0.001$) for all ensemble configurations with more than one node, that is, in all the cases in which decision fusion techniques are applied.

VI. CONCLUSION

In this work we have proposed a novel method for taking advantage of spatial information to aid multi-channel acoustic classification of weapons. The obtained results highlight the relevance of spatial diversity in an application where spatial dependence on the signals is the biggest problem. They show how classifier fusion can be an efficient strategy for audio event classification in WASN even when array processing is not feasible due to technical limitations. Furthermore the presented Maximum Likelihood fusion rule improves the performance of classic methodologies. However, in order to take advantage of this fusion rule, the classifier ensemble needs some degree of diversity, since it becomes equivalent to the average fusion rule when a single covariance matrix is considered throughout the ensemble. On top of this, covariance matrix estimations can pose a problem in scenarios where the available data is scarce thus, it needs to be approached with care. We have also shown that spatial information retrieval from single channel gunshot recordings using pattern recognition techniques is a feasible option, specially when using an adequate feature set tailored to the particularities of the scenario, and how D&C strategies can be applied to simplify the complexity of the problem. Our LS-LDA tree yields better results than some renowned and widely used classifiers while having a lower computational complexity. The proposed methodology can be exploited with a larger number of weapon classes. For as long as a given weapon class encompasses weapons with similar and characteristic features, such as barrel length, caliber, propeller load, projectile velocity, etc, it would be possible to differentiate those weapons as belonging to a particular group.

References

[1] P. G. Weissler and M. T. Kobal, "Noise of police firearms," *J. Acoust. Soc. Amer.*, vol. 56, no. 5, pp. 1515–1522, 1974.

[2] K. S. Fansler, W. P. Thompson, J. S. Carnahan, and B. J. Patton, "A parametric investigation of muzzle blast," DTIC Document, Defense Technical Information Center, Fort Belvoir, VA, USA, Tech. Rep. ARL-TR-227, 1993.

[3] R. C. Maher, "Acoustical characterization of gunshots," in *Proc. IEEE Sig. Process. Appl. Public Security Forensics.,* Apr. 2007, pp. 109–113.

[4] A. Kawalec, J. Pietrasiński , and E. Danicki, "Selected problems of sniper acoustic localization," in *Proc. Battlefield Acoust. Sensing ISR, RTO-MP-SET-107*, 2006, paper no. 23, pp. 1–8.

[5] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, 2002.

[6] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 343–356, Feb. 2013.

[7] 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 Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1792–1802, Dec. 2014.

[8] S. Khan, A. Divakaran, and H. S. Sawhney, "Weapon identification using hierarchical classification of acoustic signatures," *Proc. SPIE*, vol. 7305, pp. 730 510–730 510, 2009.

[9] J. Sallai, W. Hedgecock, P. Volgyesi, A. Nadas, G. Balogh, and A. Ledeczi, "Weapon classification and shooter localization using distributed multichannel acoustic sensors," *J. Syst. Archit.*, vol. 57, no. 10, pp. 869–885, 2011.

[10] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proc. IEEE Int. Conf. Multimedia Expo*. 2005, pp. 1306–1309.

[11] I. L. Freire and J. A. Apolinário Jr., "Gunshot detection in noisy environments," in *Proc. 7th Int. Telecommun. Symp.*, Manaus, Brazil, 2010, pp. 1–4.

[12] T. Ahmed, M. Uppal, and A. Muhammad, "Improving efficiency and reliability of gunshot detection systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 513–517.

[13] J. Millet and B. Baligand, "Latest achievements in gunfire detection systems," in *Proc. Battlefield Acoust. Sensing ISR, RTO-MP-SET-107*, 2006, paper no. 26, pp. 1–4.

[14] S. Khanal, H. F. Silverman, and R. R. Shakya, "A free-source method (FrSM) for calibrating a large-aperture microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 8, pp. 1632–1639, Aug. 2013.

[15] D. Ayllón, H. A. Sánchez-Hevia, R. Gil-Pita, M. U. Manso, and M. R. Zurera, "Indoor blind localization of smartphones by means of sensor data fusion," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 4, pp. 783–794, Apr. 2016.

[16] R. C. Maher and S. R. Shaw, "Directional aspects of forensic gunshot recordings," in *Proc. Audio Eng. Soc. Conf., 39th Int. Conf., Audio Forensics. Practices Challenges*, Jun. 2010, paper no. 4-2.

[17] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *Proc. IEEE 18th Symp. Commun.. Veh. Technol. Benelux.*, 2011, pp. 1–6.

[18] J. H. Kotecha, V. Ramachandran, and A. M. Sayeed, "Distributed multitarget classification in wireless sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 703–713, Apr. 2005.

[19] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[20] L. Lam and C. Y. Suen, "Application of majority voting to pattern recognition: an analysis of its behavior and performance," *IEEE Trans. Syst., Man, Cybern., Part A, Syst. Humans*, vol. 27, no. 5, pp. 553–568, Sep. 1997.

[21] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *Proc. IEEE, 22nd Eur. Signal Process. Conf.,* 2014, pp. 2375–2379.

[22] H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins, "A multi-channel fusion framework for audio event detection," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.,* 2015, pp. 1–5.

[23] S. Gergen, A. Nagathil, and R. Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Process.*, vol. 107, pp. 21–32, 2015.

[24] B. Chen, R. Jiang, T. Kasetkasem, and P. K. Varshney, "Channel aware decision fusion in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 52, no. 12, pp. 3454–3458, Dec. 2004.

[25] T.-Y. Wang, L.-Y. Chang, D.-R. Duh, and J.-Y. Wu, "Fault-tolerant decision fusion via collaborative sensor fault detection in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 756–768, Feb. 2008.

[26] D. Ciuonzo, G. Romano, and P. S. Rossi, "Channel-aware decision fusion in distributed mimo wireless sensor networks: Decode-and-fuse vs. decode-then-fuse," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2976–2985, Aug. 2012.

[27] M. Duarte and Y.-H. Hu, "Distance-based decision fusion in a distributed wireless sensor network," *Telecommun. Syst.*, vol. 26, no. 2–4, pp. 339–350, 2004.

[28] H. A. Sanchez-Hevia, D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera, "Gunshot classification from single-channel audio recordings using a divide and conquer approach." in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2015, pp. 233–240.

[29] H. Parvin, H. Alinejad-Rokny, and S. Parvin, "Divide and conquer classification," *Australian J. Basic Appl. Sci.*, vol. 5, no. 12, pp. 2446–2452, 2011.

[30] J. C. Freytag, D. R. Begault, and C. A. Peltier, "The acoustics of gunfire," in *Proc. Int. Congress Expo. Noise Control Eng.*, vol 3, 2006, pp. 1470–1480.

[31] *Acoustics. Noise From Shooting Ranges. Part 2: Estimation of Muzzle Blast and Projectile Sound by Calculation*, ISO-CEN.17201-2, 2006.

[32] W. B. Karl Wilhem Hirsch, "Estimation of the directivity pattern of muzzle blasts," in *Proc. AIA-DAGA*, 2013, pp. 961–963.

[33] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. New York, NY, USA: Wiley, 2012.

[34] M. Hunt, M. Lennig, and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," in *Proc. IEEE Int. Conf Acoust., Speech, Signal Process.*, 1980, vol. 5, pp. 880–883.

[35] M. A. Figueiredo, J. B. Dias, J. P. Oliveira, and R. D. Nowak, "On total variation denoising: A new majorization-minimization algorithm and an experimental comparisonwith wavalet denoising," in *Proc. 2006 IEEE Int. Conf. Image Process.*, 2006, pp. 2633–2636.

[36] J. Ye, "Least squares linear discriminant analysis," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1087–1093.

[37] A. R. Abu-El-Quran, R. Goubran, and A. D. Chan, "Security monitoring using microphone arrays and audio classification," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 4, pp. 1025–1032, Aug. 2006.

[38] U. Srinivas, N. M. Nasrabadi, and V. Monga, "Graph-based multi-sensor fusion for acoustic signal classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 261–265.

[39] D. M. Tax, R. P. Duin, and M. Van Breukelen, "Comparison between product and mean classifier combination rules," in *Proc. Workshop Statist. Pattern Recognit.*, Prague, Czech Republic, 1997, pp. 165–170.

[40] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, 1994.

[41] B. Malhotra, I. Nikolaidis, and J. Harms, "Distributed classification of acoustic targets in wireless audio-sensor networks," *Comput. Netw.*, vol. 52, no. 13, pp. 2582–2593, 2008.

[42] J. Szurley, A. Bertrand, P. Ruckebusch, I. Moerman, and M. Moonen, "Greedy distributed node selection for node-specific signal estimation in wireless sensor networks," *Signal Process.*, vol. 94, pp. 57–73, 2014.

[43] H. Gish, “A probabilistic approach to the understanding and training of neural network classifiers,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 1990, pp. 1361–1364.

[44] BOOM Library GbR, “GUNS—Construction Kit,” 2011. [Online]. Available: http://www.boomlibrary.com/boomlibrary/products/guns

[45] B. Efron, “Bootstrap methods: Another look at the jackknife,” *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, Jan. 1979.

[46] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011.

Héctor A. Sánchez-Hevia (S’14) was born in Asturias, Spain. He received the B.Eng. degree in electrical and electronic engineering from the University of Oviedo, Asturias, Spain, in 2009, and the M.Sc. degree in electronic engineering and the M.Sc. degree in information and communication technologies, in 2013 and 2015, respectively, from the University of Alcalá, Madrid, Spain, where he is currently working toward the Ph.D. degree. His current research is focused on array signal processing and machine learning applications for Audio, especially wireless audio sensor networks.

David Ayllón (S’11–M’13) received the B.Sc. (Hons.) degree in telecommunication engineering from the University of Valladolid, Valladolid, Spain, in 2006, the M.Sc. (Hons.) degree in biomedical engineering from the University of Boras, Boras, Sweden, in 2009, and the M.Sc. and Ph.D. (Hons.) degrees in information and communications technologies from the University of Alcalá, Madrid, Spain, in 2009 and 2013, respectively. His dissertation was on speech enhancement algorithms for audiological applications. He is currently working as a Machine Learning Scientist at Fonetic, Madrid, Spain. His current research interests include machine learning for audio and text processing, speech recognition, and natural language processing.

Roberto Gil-Pita (S’02–A’05–M’09) received the M.Eng. degree in telecommunication engineering and the Ph.D. degree (Hons.) in electrical engineering from the University of Alcalá, Madrid, Spain, in 2001 and 2006, respectively. From 2001, he has worked at the Department of Signal Theory and Communications, University of Alcal, in the Applied Signal Processing Research Group. His research interests include pattern recognition and audio signal processing, focusing on sound source separation, hearing aids, and emotional speech. In these fields, he is author of more than 30 journal papers included in the Journal Citation Report, and around 90 conference papers. He is also Project Manager of several projects with public and private fundings, including the 2-year ATREC project for the real-time analysis of combat stress, funded by the Spanish Ministry of Defense, and the SSPressing-Colist project for smart audio processing, funded by the Spanish Ministry of Economy and Competitiveness.

Manuel Rosa-Zurera (SM’08) received the B.Eng. (Hons.) degree in technical telecommunication engineering from the University of Alcalá, Madrid, Spain, in 1990, the M.Eng. degree in telecommunication engineering from the Technical University of Madrid, Spain, in 1995, and the Ph.D. (Hons.) degree from the University of Alcalá, Madrid, Spain, in 1998. From 1991 to 1997, he worked as a Researcher and Lecturer at the Department of Circuits and Systems Engineering, Technical University of Madrid. Since 1997, he has worked at the Department of Signal Theory and Communications, University of Alcalá, where he is a Full Professor since 2010. He has been the Head of the department from 2004 to 2010, and the Dean of the Polytechnic School from 2010 to 2017 . His research interests include statistical signal processing, signal models, source coding, speech and audio signal processing, and radar signal processing, areas in which he has been involved in many research projects, and has published more than 50 papers in international journals.