TU Delft

Delft University of Technology

Urban Objects Classification With an Experimental Acoustic Sensor Network

de Groot, Teun H.; Yarovoy, Alexander G.; Woudenberg, E

DOI
10.1109/JSEN.2014.2387573

Publication date
2015

Document Version
Accepted author manuscript

Published in
IEEE Sensors Journal

Citation (APA)
de Groot, T. H., Yarovoy, A. G., & Woudenberg, E. (2015). Urban Objects Classification With an Experimental Acoustic Sensor Network. *IEEE Sensors Journal*, *15*(5), 3068-3075. https://doi.org/10.1109/JSEN.2014.2387573

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSEN.2014.2387573

Urban Objects Classification with an Experimental Acoustic Sensor Network

Teun H. de Groot, Evert Woudenberg, and Alexander G. Yarovoy, *IEEE Fellow*

***Abstract*—This paper proposes feature extraction methods for object classification with passive acoustic sensor networks deployed in (sub-)urban environments. We analyzed the emitted acoustic signals of three object classes: guns (muzzle blast), vehicles (running piston engine) and pedestrians (several footsteps). Based on the conducted analysis, methods are developed to extract features that are related to the physical nature of the objects. In addition, localization methods are developed (e.g. pseudo-matched-filter), because the object location is required for one of the feature extraction methods. As a result, we developed a proof-of-concept system to record and extract discriminative acoustic features. The performance of the features and the final classification are assessed with real measured data of the three object classes within sub-urban environment.**

***Index Terms*—classification, localization, acoustic sensors, urban environment.**

I. Introduction

SURVEILLANCE in urban setting has attracted noticeable attention and many sensor types exist that can contribute to this purpose. A new trend it that the required surveillance is not limited anymore to traditional detection and/or localization, but that it is extended to object classification. Classification is required to support decision-making (e.g. is the object unusual/threatening?). As traditional sensors do not always provide sufficient means to solve this problem on their own (e.g. electro-optical devices during fog), passive acoustic sensors are seen as an alternative source of information.

The physical limits and challenges for acoustic surveillance depend on acoustic wave propagation. This propagation is investigated in many studies (e.g. [1], [2]), and some even acuminate their study on urban environments (e.g. [3], [4]). A few example physical effects are [5], [6]: reflections, scattering, and diffraction. As some of these effects depend on the frequency and the medium, the received acoustic signal depends on the source's position and on its sound spectrum. Moreover, the signals can be hampered and altered by many environmental factors such as interference by wind, precipitation and other noises.

There are a number of operational advantages with acoustic surveillance: (i) sound cannot be easily damped or blocked, which makes it hard for an object to stay undetected; and (ii) acoustic sensors do not have to transmit signals and the system can remain hidden. However, there also remain operational challenges: (i) objects are usually non-cooperative, and their transmitted acoustic signals are not optimized for classification purposes; and (ii) acoustic sensors can potentially be "fooled" by artificial sounds. Therefore, the use of acoustic sensors in mission-critical applications is limited. However, adding them to other types of sensors (e.g. radar systems) within a heterogeneous sensor network can improve the overall surveillance.

Many current feature extraction methods for acoustic surveillance are limited to processing techniques such as the Fourier transform [7], [8], [9], Wavelet analysis [10], [11], [9] and Cepstrum [8], [12], [13]. Usually, the recorded signals are processed with these techniques and the output is given to a well-known classification method. Next, the processing techniques or the classification methods are compared with each other based on the classification performance (e.g. [14], [15]). The most frequently used method relies on neural networks to automatically extract the features (e.g. [7], [12], [16]). Although above mathematical techniques require little design efforts, they do not provide insight into the relevant features that are actually related to the nature of the object.

This paper proposes object features that can be extracted from acoustic recordings and used to discriminate between object classes. We focus in this study on three object types, which all produce their own specific sounds: guns, vehicles and pedestrians. A challenge is that some sound sources are not acoustically consistent (also discussed in [17]) and the signals can still differ significantly within the same object class. To overcome this, we propose features that are related to the object's nature and physically explainable. The methods are tested within a proof-of-concept system that recorded real data. As a result, we also assess the surveillance performance of a network of passive microphones to classify (and localize) objects in urban environment.

The structure of this paper is as follows. The sound signals of the selected objects (i.e. gun muzzle blasts, running piston engines and walking pedestrians) are analyzed in Section II. With the gained knowledge, methods are presented in Section III to extract object features that can be used for classification. Section IV presents methods to perform object localization, because the gunshot feature extraction requires the object distance relative to the sensors. The resulting performance of the feature extraction process is estimated within a proof-of-concept system in Section V. Section VI summarizes the results and concludes this study.

T. H. de Groot is with the Delft University of Technology, 2628 CD, The Netherlands (e-mail: T.H.deGroot@tudelft.nl; phone: +31 (0)15 27 87089).

E. Woudenberg is with Thales Nederland B.V., Hengelo, 7554 PA, The Netherlands (e-mail: evert.woudenberg@nl.thalesgroup.com).

A. Yarovoy is with the Delft University of Technology, 2628 CD, The Netherlands (e-mail: A.Yarovoy@tudelft.nl).

Manuscript received April 09, 2014; revised December 18, 2014.

Copyright (c) 2014 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSEN.2014.2387573

II. Acoustic Analysis

As recorded signals depend on the used hardware, the recording hardware is first briefly discussed. Next, we analyze the acoustic signals of gun muzzle blasts, running piston engines and walking pedestrians. The findings that are relevant for feature extraction are summarized after each class analysis.

A. Recording hardware

Four microphones and recording devices were available for the proof-of-concept system. The FG-3329 microphones that are omni-directional according to their specifications from Knowles electronics were each attached to their own amplifier. The amplifiers amplified with a factor equal to 20 up to 48kHz with internal noise proportional to 1/f. We used a cap and windjammer from Raycote Lavaliere. The cap is made from foam and the jammer from fur, which is an artificial fiber fixed to a fabric mesh backing. The digital sound card had a maximum sample frequency of 96 kHz with 24 bit and dynamic range of 99dB.

B. Muzzle blast

A conventional firearm uses an explosive charge to propel the bullet out of the gun barrel. The resulting sound is called a gunshot [18], [19]. We recorded many gunshot signals at two shooting clubs with different guns, calibers and backgrounds noises. The first club was indoor and resulted in too much reverberation effects for an in-depth analysis to present in this paper, but the second one was outdoor that resulted in better recordings for analysis.

Fig. 1 shows an example gunshot signal of a shotgun caliber 12 under/over. The Line-of-Sight (LOS) and a reflection signal are received around 0.01 and 0.02 seconds respectively. We were surprised of the loud signals at a range of 10 meters: the used hardware was not able to cope - even after placing an attenuator over the microphone - as can be seen by an artifact (i.e. straight 45 angle line) around 0.01 seconds. This artifact can be identified as clipping in the amplifier. When the microphone was placed in front of the gun the received power further increased, but when it was placed behind the shooter at 10 meters the signal in Fig. 2 was recorded without any artifact. The spectrogram shows that the bandwidth of the gunshot goes up to 48 kHz.

Four main gunshot features are observed during the measurements. Firstly, the emitted energy is extremely high, which resulted sometimes in clipping. Secondly, the received LOS signal is angle dependent: received power in front of the gun is the strongest and behind the gun the lowest. A similar conclusion on angle dependency was made in [19]. Thirdly, the spectrogram shows that the signal is short in time, but wide-band in frequency including ultrasound frequencies up to 48 kHz. Fourthly, with the same ammunition and angle, the gunshot sounds are very consistent and identical.

C. Running piston engine

Vehicle acoustic signatures depend on the type and dynamics such as engine speed, load and road surface [10]. The

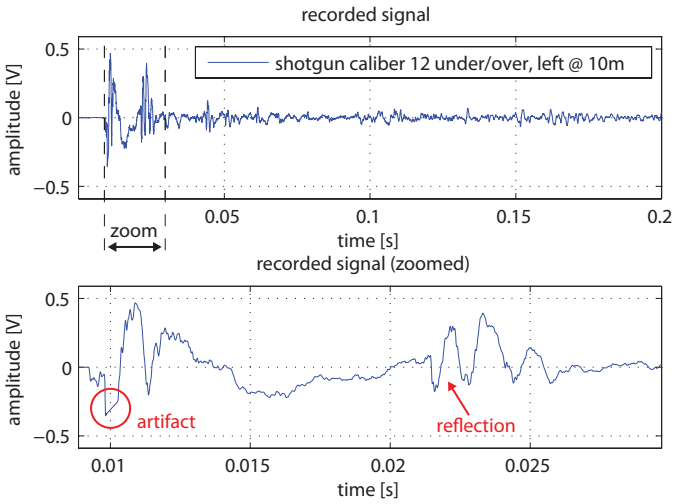


Fig. 1. Shotgun signal recorded from the left side at 10 meter (with artifact).

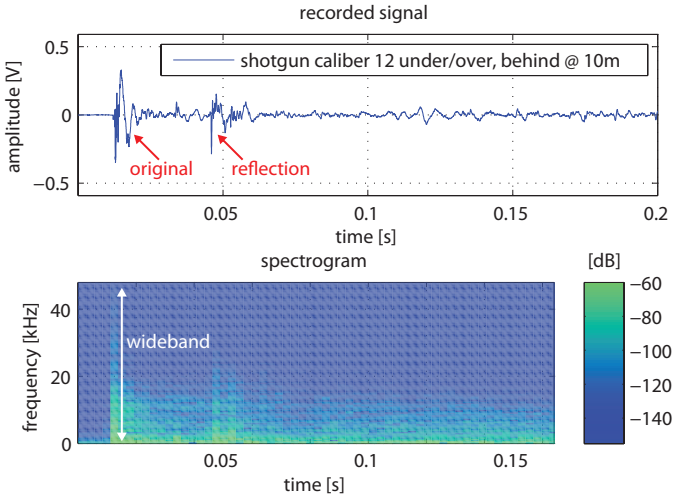


Fig. 2. Shotgun signal recorded from behind at 10 meter (without artifact).

sounds are mainly caused by [8]: rotational parts, vibrations in the engine, friction between the tires and the pavement, wind effects, gears and fans. The vehicle sound spectrum can be analyzed as in [9] and it is suggested by [7] that the main vehicle features are within the low frequencies.

Acoustic measurements were taken at different distances from a running piston engine (four-cylinder piston in Renault Laguna 2.2 liter and Honda Civic Shuttle). The initial measurements showed that the engine signal is wide-band with frequencies up to 40 kHz, and that most of the power is located below 10 kHz. The interesting features are located below 400 Hz: Fig. 3 shows a spectrogram of a recording where the driver (with four cylinders engine) continuously varied the engine speed. As can be seen, the signal harmonics are changing over time. Note, also the emitted power varies depending on the engine speed.

It is known that every cylinder of a four-stroke engine sparks every two revolutions. Thus, two explosions occur at every revolution with four cylinders. For example, with 1000 Revolutions Per Minute (RPM) and two explosions every

Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSEN.2014.2387573

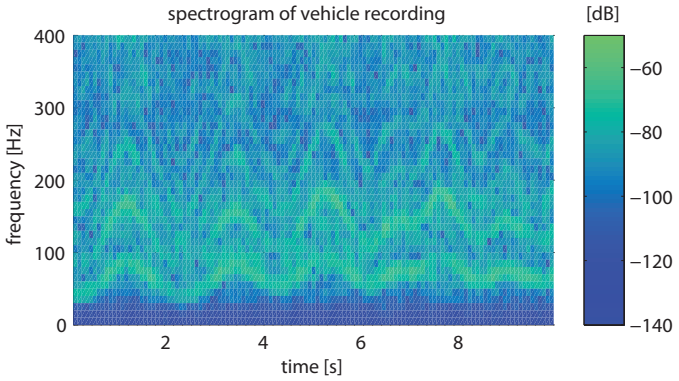


Fig. 3. Spectrogram of piston engine that is varying speed.

revolution, the explosion frequency is around 33 Hz. In fact, this fundamental frequency is found in our recordings, for example, in Fig. 3. Beside this feature, no distinguishable sharp time features were found which could be extracted (e.g. we also thought about gearbox sounds). Therefore, it is hard to unambiguously estimate the beginning (or ending) of the received signal (e.g. for time-based localization).

Altogether, the following three conclusions were made. Firstly, the emitted engine signal is a wide-band signal including ultrasound frequencies, and transmitted power depends on the engine speed. Secondly, no significant signal difference was observed when the recording angle was changed: only some little attenuation, but that is caused by the vehicle components and distance between the engine and microphone. For example, we did not find any special exhaust signals. Thirdly, the fundamental harmonic frequency is, as explained above, understandable linked to the engine speed.

D. Walking pedestrian

The sound of a person's footstep is determined by three dominant conditions [20], [11]: footwear (e.g. sneakers, bare foot), ground surface (e.g. concrete, wood) and gait (e.g. personal motion, speed). The main difficulty for feature extraction is the change in these three conditions. Although even identification based on footsteps is suggested [11], during our measurements it became clear that the signal depends too much on the footwear and that the received power is very low for footstep classification even at small distances (e.g. 2 meters). Fig. 4 shows an example recording of a few footsteps at 1 meter. As can be seen, it is barely observable.

To improve the footstep sound analysis we selected the best ground, shoes and environment (indoor) to increase transmit power and decrease environmental noise. Fig. 5 shows the best recorded signal. In this recording two separate signals were received. First from the heel bone and second from the metatarsus which is located between the toes and mid-foot. Value $t_{footstep}$ is the time between these two signals. We also discovered that the left and right footstep sounds are not identical. Furthermore, a left or right footstep is not necessary identical to its previous one.

Four conclusions are derived from the footsteps analysis. Firstly, the signal strongly depends on gait, shoes and ground:

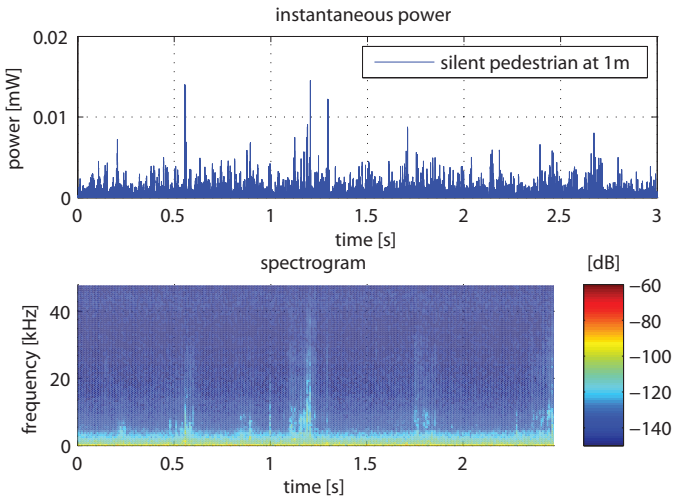


Fig. 4. Instantaneous power and spectrogram of multiple poor footsteps.

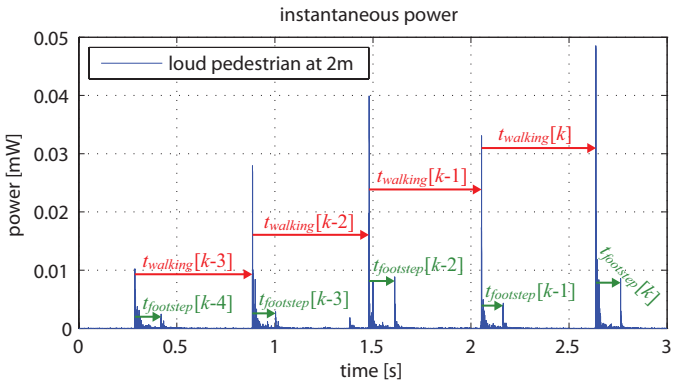


Fig. 5. Example of time differences of a walking pedestrian.

each composition produces a unique sound. Moreover, even with the same composition, the footstep signals are unpredictable and not consistent. The practical use of individual footstep sounds is therefore limited for classification. Secondly, the striking force signal is the strongest and wide-band. Depending on composition, ultrasound frequencies up to 40 kHz are present due to the striking and friction. Thirdly, the received signal power is low. Poor footsteps are barely detectable at 1 meter, but also good footsteps at 6 meters are problematic to detect. Fourthly, walking is a periodic activity and the time between two footsteps, which can be used as a feature, does not vary that much among the pedestrians.

III. Feature Extraction

The key behind successful object classification is to construct signatures that are sufficiently different between objects of different classes, and, at the same time, that they are similar for objects of the same class. In order to build such signature, a specialized feature extractor is developed for each object type. The combination of all extracted features compose the final acoustic signature.

Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSEN.2014.2387573

TABLE I
FEATURE EXTRACTION FOR RUNNING ENGINE

Number	Feature	Expected value
E1	Harmonic	$25 \leq f_0 \leq 120$
E2	Peak ratio 1	$1.9 \leq \frac{f_1}{f_0} \leq 2.1$
E3	Peak ratio 2	$2.9 \leq \frac{f_2}{f_0} \leq 3.1$
E4	Harmonic derivative	$0 \leq \left\lvert \frac{d}{dt} f_0 \right\rvert \leq 40$

TABLE II
FEATURE EXTRACTION FOR PEDESTRIAN

Number	Feature	Expected value
P1	Walking time	$0.45 \leq t_{walking}[k] \leq 1.0$
P2	Walking ratio	$0.7 \leq \frac{t_{walking}[k]}{t_{walking}[k-1]} \leq 1.3$
P3	Footstep time	$0.05 \leq t_{walking}[k] \leq 0.2$
P4	Footstep ratio	$0.8 \leq \frac{t_{walking}[k]}{t_{walking}[k-1]} \leq 1.2$

A. Extracting Emitted-Spectrum

The exact emitted gunshot sounds differ over the different gun types. However, the transmitted power-spectrum could be a good feature to classify a gunshot (i.e. muzzle blast), because it can show if the emitted power is high enough and if the signal is wide-band. The transmitted spectrum can be reconstructed based on the received spectrum, object distance and propagation loss model:

$$P_0(f) = \frac{P_n(f)}{G_n(f)A_\alpha(d_n, f)A(d_n)} \tag{1}$$

where $P_0(f)$ is the transmitted power as a function of frequency f, $P_n(f)$ is the received power at microphone n, $G_n(f)$ is the gain of microphone n, $A_\alpha(d_n, f)$ the sound absorption function, $A(d_n)$ the attenuation function due to geometric spreading of sound and d_n the distance between the object and microphone n.

The object position is required for this feature extraction, and therefore, object localization is presented in Section IV. Note that it is not known which frequencies are emitted. Therefore, a potential drawback of this approach is that noises, which were not emitted by the object, are also amplified and considered part of the emitted-spectrum.

B. Extracting Harmonic-Formation

A large variety of vehicles exist, but eventually it is just a vibrating object due to a running engine with periodic explosions. Therefore, the signal harmonics can be used for classification. We derived the four features of Table I.

The fundamental frequency f_0, which is directly linked to the engine speed, is the first engine feature. However, one peak within the spectrum range of roughly 20Hz to 150Hz is only enough for an indication. The second and third features appoint the constellation of the multiple spectrum peaks and is needed to confirm that indeed the peaks are multiple harmonics (e.g. ratio between the frequencies f_0, f_1 and f_2). The last feature is the derivative of the fundamental frequency and can be used to confirm that a driver is changing the RPM.

C. Extracting Time-Intervals

The variety of sounds created by the footstep force or friction is in practice too large for one-footstep recognition. However, detection of walking is possible and less dependent on the exact composition of shoes, ground and gait. Fig. 5 shows useful time-intervals that can be extracted. The chosen features are outlined in Table II.

The walking time $t_{walking}$ is the first pedestrian feature. The second feature is the ratio between the last two walking times. The third feature is the footstep times $t_{footstep}$ and the fourth feature the ratio between the last two footstep times. Due to the fact that the second peak within a footstep is frequently not present, the last two features are mainly intended for confirmation. The first and third features dependent on a particular gait, but the second and fourth features are less dependent on this and indicating the continuity of the walk (e.g. not injured).

IV. OBJECT LOCALIZATION

The gunshot feature extraction requires the object position. There are two methods available for passive position estimation: time-based localization using the propagation time model (i.e. Time Difference Of Arrival), and power-based localization using the propagation loss model (i.e. Received Signal Strength). This section presents techniques to extract the required information for time-based localization purposes.

A. Time of Arrival

The geographical object position can be estimated based on the arrival times of the LOS signals at the nodes, because the propagation time model is a function of object-node distance:

$$TOA_n = TOE + \frac{d_n}{c_{air}} + TE_n \tag{2}$$

where TOA_n is the time of arrival at node n, TOE the time of emission, d_n the distance between the object and node n, c_{air} the speed of sound in air and TE_n the time error for node n due to modeling error and observation noise. The three object position coordinates and one TOE can be estimated by using minimal four TOA_n and minimizing the value TE_n.

B. Pseudo-matched-filter

A matched-filter, based on the original transmitted signal, would be ideal for estimating TOA_n to get a high time accuracy. However, the transmitted signal is not available. Therefore, a mask is extracted 'on-the-run' from one of the recorded signals to filter (i.e. cross-correlation) all the other received signals. This 'pseudo-matched-filter' method is comparable with other techniques for TDOA (e.g. [21], [22]),

Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSEN.2014.2387573

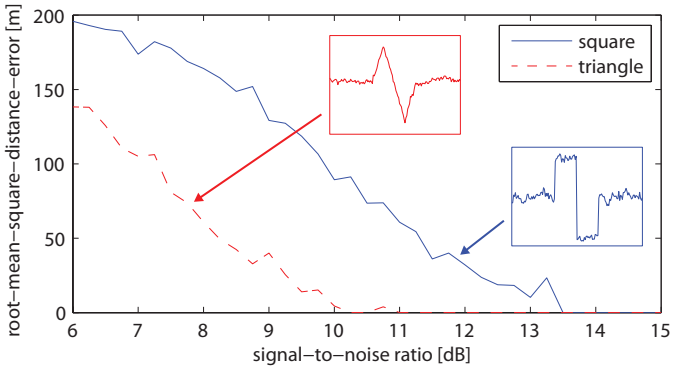


Fig. 6. Simulated performance of pseudo-matched-filter.

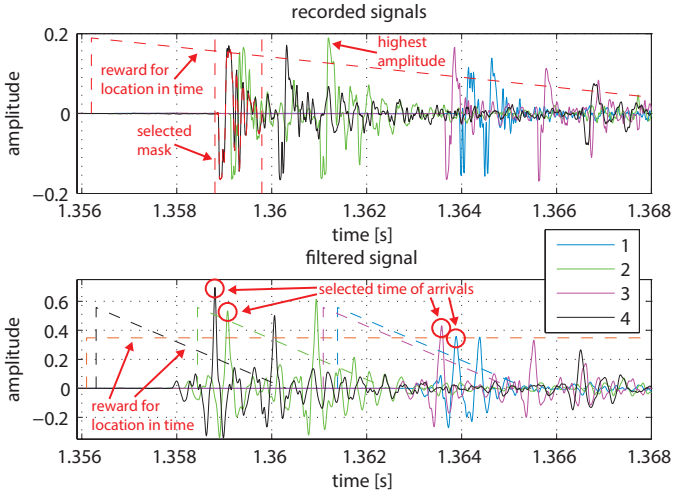


Fig. 7. Arrival time extraction for time-based localization.

but a difference is that only a small part of the signal is selected to filter the much longer signals.

We simulated the pseudo-matched-filter approach to estimate its performance. Assume one clean signals s based on a square or triangle shape. Next, two input signals y_1, y_2 are created by adding (pink) noise signals n_1, n_2 to the signal s: $y_1 = s + n_1$ and $y_2 = s + n_2$. An appropriate filter-mask is constructed based on a searched within signal y_1. This mask is used to filter signal y_2, and then, the time-of-arrival is estimated and compared with the actual time-of-arrival. The averaged distance error (i.e. time error TE multiplied with speed of sound c_{air}) as functions of Signal-to-Noise-Ratio (SNR) is given in Fig. 6. Although multi-path (i.e. multiple pulses in y_1 and y_2) and signal alteration (e.g. sound absorption as function of frequency) is not considered in this analysis, it gives an initial performance indication.

C. Rejection of non-LOS signals

In order to extract the correct TOA_n values, the following has to be considered: the suppression of non-LOS signals. A trivial strategy that selects the signal with the highest amplitude is not a successful one, because the LOS signals did not always have the highest amplitude during our experiments. Therefore, a strategy based on rating is implemented.

Fig. 7 shows an example of time of arrival extraction with four received signals using the pseudo-matched-filter. The upper plot shows the recorded signals and the mask-selection. As can be seen, the signal which was selected for the mask did not have the highest peak (signal two has the highest peak), but this signal was received earlier which indicates less mutations. The mask-selection method is based on the principle to give signal peaks points for their time properties (illustrated by the sloped dotted lines) and their amplitude:

$$k_s = \arg\max_k \left(W_a\left[y_p(k)\right] + W_b\left[t_p(k) - t_p(k=1)\right]\right) \quad (3)$$

where W_a and W_b are the weight functions to rate the amplitude $y_p(k)$ and moment $t_p(k)$ of peak k respectively. $y_p(k)$ and $t_p(k)$ include all extracted peaks of all the nodes ordered in time. The area around the peak with the highest score is selected for the mask to filter all the recorded signals.

A similar approach (e.g. [23]) is used for arrival time estimation in the filtered signals as shown in the lower plot of Fig. 8. Again, the peaks receive points for their time and amplitude properties:

$$TOA_n = t_n(\arg\max_k W_1\left[y_n(k)\right] + W_2\left[t_n(k) - t_n(1)\right] + W_3\left[t_n(k) - t_p(k_s)\right]) \quad (4)$$

where W_1, W_2 and W_3 are the weight functions (illustrated by dotted curves) for peak k. $y_n(k)$ and $t_n(k)$ are the extracted peaks from the filtered signal n. The time of the peak with the highest score is used to estimate TOA_n.

Above rating-method is flexible, able to cope with many types of incoming signals, and was successful in selecting the LOS signals and ignoring the reflections. With a SNR of 15dB the method was always successful in our experiments. The method sometimes works with lower SNR (e.g. with very silent footsteps), but the probability of incorrect positions will increase and the position error will probably not be Gaussian.

V. Experimental Evaluation

A proof-of-concept system is developed to investigate the surveillance performance of an acoustic sensor network in (sub-)urban environments. The experimental setup with the hardware as discussed in Section II is shown in Fig. 8. Note that the microphones are mounted on different heights to robustly estimate three-dimensional positions and extract features from different angles. We recorded 90 seconds without objects, 42 (toy-gun)shots of 3 seconds, 12 walks that lasted 12 seconds of 6 different pedestrians and 200 seconds of 2 different vehicles. Each recording was partitioned into 0.5 seconds long recordings that have a 10% overlap with their previous recording. The classifier should determine if an object(-class) is presented in the environment based on such a partitioned recording, and it can use information that is saved during the analysis of the previous recordings (e.g. a pedestrian walk takes longer than 0.5 seconds). The signal processing is implemented with MATLAB.

Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSEN.2014.2387573

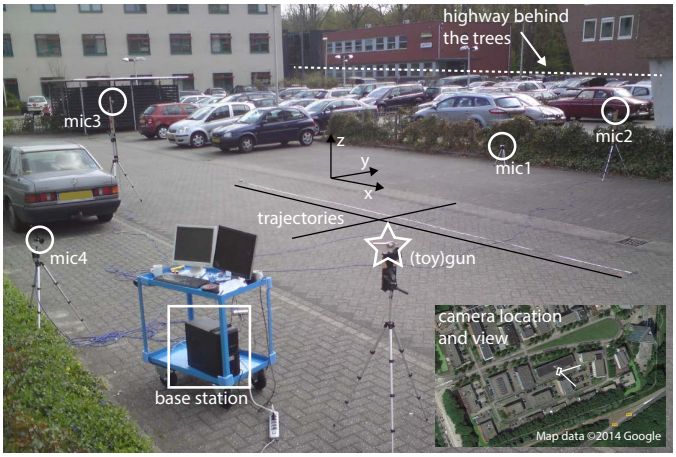


Fig. 8. Measurement setup and experimental system at a parking area.

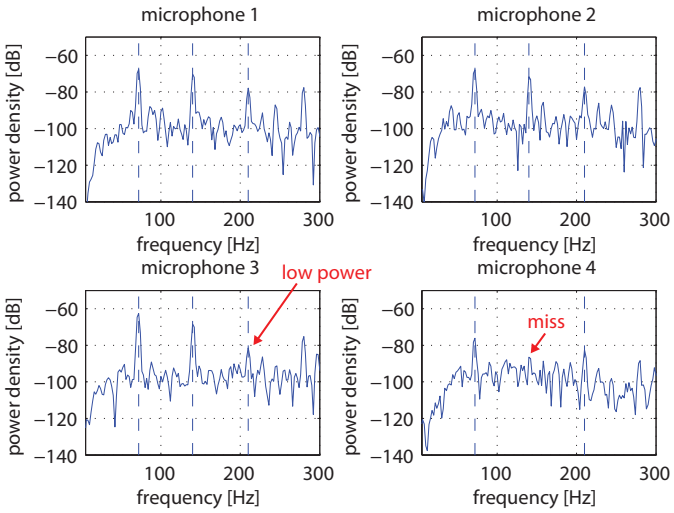


Fig. 9. Parallel feature extraction for vehicle classification.

A. Localization performance

A non-linear least square estimator is used to estimate the object position based on the extracted information and propagation models. The Root Mean Square Error (RMSE) and Standard Deviation (SD) performance is measured with toy-gunshots. A toy-gunshot provides a similar sound as a real gunshot, but its emitted energy is lower and the exact signal varies (signals of real gunshots do not vary noticeable with the same gun, ammunition and recording angle). Thus, localizing a toy-gunshot is slightly more difficult than localizing a real (unsilenced-)gunshot, but the retrieved SNR of the sounds was still above 15dB at each microphone. The time-based localization (based on a 1 millisecond sized mask in the pseudo-matched-filter) was quite accurate: $RMSE_{3D}$ = 0.29m and SD_{3D} = 0.04m. The difference between the RMSE and SD even indicates that there is room for performance improvements (e.g. calibrate node positions).

B. Advantage of sensor network

The system consists of several sensors. To benefit from the spatially distributed microphones, a fusion strategy is used for each feature extraction. The final emitted-spectrum feature is constructed by taking the mean spectrum of the multiple emitted-spectra, that were estimated with the multiple microphones. The final harmonic-formation features are estimated by the weighted mean of the extracted engine features at each microphone, where the weights are proportional to the received power. The final time-intervals are estimated by constructing the peak information based on one microphone that recorded the signal with the highest received power.

Although more sophisticated solutions exist for information fusion, the above methods already make the feature extraction more robust. For instance, the footstep time-interval extraction may require only one microphone, but the difficulty is that two/three good footsteps in a row are required. We experienced that when footsteps of the same walk were received by different sensors, and their information was fused, the object was still successfully classified as a pedestrian.

Another example can be given related to engine detection. A vehicle recording of four nodes is shown in Fig. 9 where it is difficult to cope with a too low SNR. As can be seen, the spectra differ from each other because the sensors recorded the vehicle from different angles and distances. As a result, the spectrum peaks were not always correctly selected in each recording. However, based on the fused information, the object could still be classified as a vehicle, which confirms the added value of sensor networks.

TABLE III
DETECTION PERFORMANCE INCLUDING FALSE POSITIVES

Detected Feature(s) \ Actually Recorded	Nothing	Gunshot	Engine	Pedestrian
G(16-24kHz)	0%	100%	0%	0%
G(0-48kHz)	0%	100%	0%	0%
E1 & E2	6%	19%	100%	9%
E1, E2 & E3	0%	7%	97%	0%
P1	0%	2%	2%	65%
P1 & P2	0%	0%	0%	35%

C. Class-detection performance

Several detectors can be developed based on the extracted features. One feature is in principle enough to construct such a detector for determining the presence of a particular object class, but the performance may not be satisfactory. Adding extra features can reduce the false positive, but it can also increase the false negative. Table III provides an overview of the detection performance of the extracted features. The two gun features G present both the total energy in the specified frequency range. The detection thresholds for the engine and pedestrian are based on the expected feature values in Table I (e.g. the feature(s) should be within the expected value range before a feature is detected). Note, no detectors are shown based on E4, P3 and P4, because the engine was not changing speed and almost all footstep sounds did not have the second footstep peak. To conclude, a detector is fully successful if it detects 100% of its own class (i.e. no false negatives) and 0% of other classes (i.e. no false positives).

Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSEN.2014.2387573

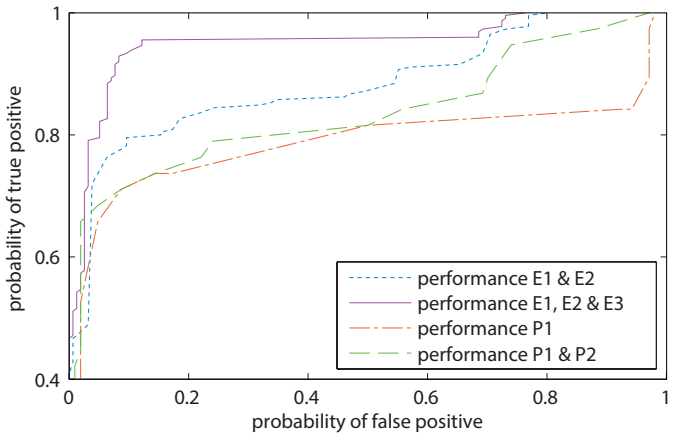


Fig. 10. Performance of extracted features for engine and pedestrian.

The (toy-)gunshot detector is very successful, because the transmitted energy of a gunshot is well separated from the ones of other classes. Vehicles can be detected with only two features, but the third feature can be add to reduce the false alarms. The pedestrian detection performs the least: the majority is only detected as pedestrian when one feature is used. Adding a second pedestrian feature reduces the detection rate significantly, because receiving three usable footstep signals in a row is in practice rare.

A more detailed performance analysis of the extracted features for engine and pedestrian can be found in Fig. 10 (i.e. receiver operating characteristic). The curves are plotted by varying the detection thresholds, which then results in different probabilities of true positives and false positives. The detection improves when the true positive rate increases for the same false positive rate or when the false positive rate decreases for the same true positive rate. In other words, when the curve moves toward the top-left corner. As can been seen, this is the case when more features are involved in the detection process.

D. Classification performance

A final classifier is constructed by combining three detectors. This classifier checks first if a gunshot is detected with the total emitted-power in the frequency band 16-24kHz. If a gunshot is not detected, then the engine detector is used based on features E1, E2 and E3. If also a vehicle is not detected, then the signal is checked for a pedestrian based only on feature P1. Thus, this tree-classifier is based on five features in total. The resulting classification performance is given in Table IV (i.e. confusion matrix). Because this simple classifier is able to provide this performance, it is certain that the discriminative information is available in the developed acoustic signature. Possibly the performance can be further improved with more sophisticated classification methods (e.g. improve detection-thresholds with machine learning).

To conclude, the proposed features, which are discriminative and physically understandable, allow a simple classifier to distinguish between the classes. Note that the SNR is crucial for feature extraction, and sometimes it was just too low to detect an object. Nevertheless, when the SNR was above 15dB,

TABLE IV
CLASSIFICATION PERFORMANCE BASED ON FIVE FEATURE INPUTS

		Actually Recorded			
		Nothing	**Gunshot**	**Engine**	**Pedestrian**
Predicted Class	**Nothing**	100%	0%	3%	35%
	Gunshot	0%	100%	0%	0%
	Engine	0%	0%	97%	0%
	Pedestrian	0%	0%	0%	65%

the emitted powers, harmonics and footstep intervals could be successful extracted in our experiments and used as features for robust classification.

VI. CONCLUSION

Feature extraction methods are developed for urban object classification based on passively recorded acoustic signals. Three different classes of object sounds are studied: gun muzzle blasts, running piston engines and walking pedestrians. Sound analysis resulted in novel features for discrimination between the classes: acoustic emitted power spectrum, acoustic harmonic formation and acoustic peak time intervals. Further, an adaptable pseudo-matched-filter is developed for object localization. All methods are evaluated with an experimental acoustic sensor network and real measured data to obtain a performance indication. We are quite satisfied with the final classification performance, because it is reasonable good considering that no sophisticated information fusion and classification techniques are used in this experiment.

Above gives an indication of the usability of the proposed features for urban objects classification. The acoustic surveillance of pedestrians has the least potential, because the SNR of most pedestrian footsteps - even after selecting them - is too low. The use of seismic sensors may result in a better SNR. In contrary, acoustic surveillance of vehicles has higher potential, but more robustness can be required for real operation (e.g. increase SNR by environmental noise reduction). Gunshot signals are received with the highest SNR, and therefore, gunshot localization is robust and its classification is feasible. Future research can focus on extending the number of objects (e.g. pedestrian together with vehicle) and adding object-classes (e.g. speech, dogs, mini-drones). Because the proposed features differ between the considered classes, simultaneous detection of multiple objects of different classes can work. It is also interesting to study new classes, their features and SNR, and relation/overlap with the proposed features.

To conclude, the operational opportunities of using solely a passive acoustic sensors for surveillance in complex urban environments is still a long way off and remains a technical challenge. However, acoustic sensors can definitely be used as an additional source of information in heterogeneous sensor networks, improve the classification process, and eventually, enhance the situational awareness.

ACKNOWLEDGMENT

The authors thank Thales Nederland B.V. for making the required equipment available and providing professional support for obtaining the presented results.

Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

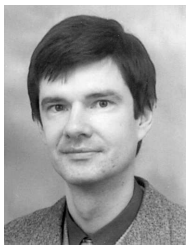
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSEN.2014.2387573

REFERENCES

[1] A. D. Pierce, *Acoustics: An Introduction to its Physical Principles and Applications*. Acoustical Society of America, 1989.
[2] J. Whitaker and K. B. Benson, *Standard Handbook of Audio and Radio Engineering*. McGraw Hill Professional, 2002.
[3] M. W. Parker, S. A. Ketcham, and H. H. Cudney, "Acoustic wave propagation in urban environments," in *DoD High Performance Computing Modernization Program Users Group Conference*, 2007, pp. 233–237.
[4] V. Singh, K. E. Knisely, S. H. Yönak, K. Grosh, and D. R. Dowling, "Non-line-of-sight sound source localization using matched-field processing," *The Journal of the Acoustical Society of America*, vol. 131, pp. 292–302, 2012.
[5] D. G. Albert and L. Liu, "The effect of buildings on acoustic pulse propagation in an urban environment," *The Journal of the Acoustical Society of America*, vol. 127, pp. 1335–1346, 2010.
[6] M. Mijić and D. Š. Pavlović, "Measurement of reverberation gain in an urban environment," *The Journal of the Acoustical Society of America*, vol. 132, pp. 1417–1426, 2012.
[7] R. H. Mgaya, S. Zein-Sabatto, A. Shirkhodaie, and W. Chen, "Vehicle identifications using acoustic sensing," in *IEEE Proceedings in SoutheastCon*, 2007, pp. 555–560.
[8] M. E. Munich, "Bayesian subspace methods for acoustic signature recognition of vehicles," in *Proceedings of the European Signal Processing Conference*, 2004, pp. 2107–2110.
[9] A. Aljaafreh and L. Dong, "An evaluation of feature extraction methods for vehicle classification based on acoustic signals," in *International Conference on Networking, Sensing and Control*, 2010, pp. 570–575.
[10] A. Averbuch, V. A. Zheludev, N. Rabin, and A. Schclar, "Wavelet-based acoustic detection of moving vehicles," *Multidimensional Systems and Signal Processing*, vol. 20, no. 1, pp. 55–80, 2009.
[11] A. Itai and H. Yasukawa, "Footstep classification using wavelet decomposition," in *International Symposium on Communications and Information Technologies*, 2007, pp. 551–556.
[12] S.-H. Shin, T. Hashimoto, and S. Hatano, "Automatic detection system for cough sounds as a symptom of abnormal health condition," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 486–493, 2009.
[13] K. Kim and H. Ko, "Discriminative training of gmm via log-likelihood ratio for abnormal acoustic event classification in vehicular environment," in *International Conference on Computers, Networks, Systems and Industrial Engineering*, 2011, pp. 348–352.
[14] M. F. Duarte and Y. Hen Hu, "Vehicle classification in distributed sensor networks," *Journal of Parallel and Distributed Computing*, vol. 64, no. 7, pp. 826–838, 2004.
[15] H. Wu and J. M. Mendel, "Classification of battlefield ground vehicles using acoustic features and fuzzy logic rule-based classifiers," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 56–72, 2007.
[16] B. A. Bucci and J. S. Vipperman, "Performance of artificial neural network-based classifiers to identify military impulse noise," *The Journal of the Acoustical Society of America*, vol. 122, pp. 1602–1610, 2007.
[17] B. Defréville, F. Pachet, C. Rosin, and P. Roy, "Automatic recognition of urban sound sources," in *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.
[18] R. C. Maher, "Acoustical characterization of gunshots," in *IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, 2007, pp. 1–5.
[19] S. D. Beck, H. Nakasone, and K. W. Marr, "Variations in recorded acoustic gunshot waveforms generated by small firearms," *The Journal of the Acoustical Society of America*, vol. 129, pp. 1748–1759, 2011.
[20] A. Ekimov and J. M. Sabatier, "Ultrasonic wave generation due to human footsteps on the ground," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 114–119, 2007.
[21] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
[22] S. Liu, C. Zhang, and Y. Huang, "Research on acoustic source localization using time difference of arrival measurements," in *International Conference on Measurement, Information and Control*, 2012, pp. 220–224.
[23] J. Scheuing and B. Yang, "Disambiguation of tdoa estimates in multipath multi-source environments (datemm)," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. 837–840.

Teun de Groot was born in the Netherlands and received his B.Sc. and M.Sc. degree in Electrical Engineering, both cum laude, in 2008 and 2010 at Delft University of Technology. His graduation project was about acoustic sensor networks and was conducted at Thales Netherlands. Currently he is a Ph.D. student at the Microwave Sensing, Signals and Systems research group of Delft University of Technology and working on resource management for reconfigurable sensors as part of the Sensor Technology Applied in Reconfigurable Systems project.

Evert Woudenberg was born in the Netherlands and received his M.Sc. degree in Applied Physics in 1999 at Delft University of Technology. In 1999 he joined Thales Nederland at the Delft facility as a radar system designer and later as advanced development engineer. In 2010 he moved to the main facility in Hengelo as portfolio development manager.



Alexander Yarovoy graduated from the Kharkov State University, Ukraine, in 1984 with the Diploma with honor in radiophysics and electronics. He received the Candidate Phys. & Math. Sci. and Doctor Phys. & Math. Sci. degrees in radiophysics in 1987 and 1994, respectively.

In 1987 he joined the Department of Radiophysics at the Kharkov State University as a Researcher and became a Professor there in 1997. From September 1994 through 1996 he was with Technical University of Ilmenau, Germany as a Visiting Researcher. Since 1999 he is with the Delft University of Technology, the Netherlands. Since 2009 he leads there a chair of Microwave Sensing, Systems and Signals. His main research interests are in ultra-wideband microwave technology and its applications (in particular, radars) and applied electromagnetics (in particular, UWB antennas). He has authored and co-authored more than 250 scientific or technical papers, four patents and fourteen book chapters. He served as a Guest Editor of five special issues of the IEEE Transactions and other journals.

Copyright (c) 2015 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.