Multimed Tools Appl (2016) 75:10407–10439
DOI 10.1007/s11042-015-3105-4

Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations

K. Lopatka[1] · J. Kotus[1] · A. Czyzewski[1]

Received: 29 August 2014 / Revised: 10 September 2015 / Accepted: 19 November 2015 /
Published online: 2 December 2015

Abstract Evaluation of sound event detection, classification and localization of hazardous acoustic events in the presence of background noise of different types and changing intensities is presented. The methods for discerning between the events being in focus and the acoustic background are introduced. The classifier, based on a Support Vector Machine algorithm, is described. The set of features and samples used for the training of the classifier are introduced. The sound source localization algorithm based on the analysis of multichannel signals from the Acoustic Vector Sensor is presented. The methods are evaluated in an experiment conducted in the anechoic chamber, in which the representative events are played together with noise of differing intensity. The results of detection, classification and localization accuracy with respect to the Signal to Noise Ratio are discussed. The results show that the recognition and localization accuracy are strongly dependent on the acoustic conditions. We also found that the engineered algorithms provide a sufficient robustness in moderately intense noise in order to be applied to practical audio-visual surveillance systems.

Keywords Sound detection · Sound source localization · Audio surveillance

1 Introduction

Recognition and localization of acoustic events are relatively recent practical applications of audio signal processing, especially in the domain of acoustic surveillance. In this case the goal is to recognize the acoustic events that may inform us of possible threats to the safety of people

✉ K. Lopatka
klopatka@sound.eti.pg.gda.pl

J. Kotus
joseph@multimed.org

A. Czyzewski
andcz@multimed.org

[1] Faculty of Electronics, Telecommunications and Informatics, Multimedia Systems Department, Gdańsk University of Technology, Gdańsk, Poland

or property. An additional information is in is the acoustic direction of arrival, which can be used to determine the position of the sound source, i.e., the place in which the event occurred.

The recognized classes of sound concerned in this work relate to dangerous events. Typically, such events include gunshots, explosions or screams [31, 42]. The majority of sound recognition algorithms described in the literature are based on the extraction of acoustic features and statistical pattern recognition [46]. Ntalampiras et al. [31] and Valenzise et al. [42] employed a set of perceptual and temporal features containing Mel-Frequency Cepstral Coefficients, Zero Crossing Rate, Linear Prediction Coefficients and a Gaussian Mixture Model (GMM) classifier. The latter work also presents sound localization techniques with a microphone array based on the calculation of the Time Difference of Arrivals (TDOA). Lu et al. [26] used a combination of temporal and spectral shape descriptors fed into a hybrid structure classifier, which is also based on GMM. Rabaoui et al. [36], Dat and Li [5] as well as Temko and Nadeau [40] proposed the utilization of Support Vector Machine classifiers (SVM) to the classification task. Dennis et al. proposed interesting methods for overlapping impulsive sound event recognition [8]. Their algorithm utilizes local spectrogram features and Hough transform to recognize the events by indentifying their keypoints in the spectrogram. A comprehensive comparison of techniques for sound recognition (including Dynamic Time Warping, Hidden Markov Models or Artificial Neural Networks) was presented by Cowling and Sitte [4]. In our approach we also propose using a threshold-based methodology for separating acoustic events from the background. A SVM classifier is used for discerning between classes of threatening events. The sound event recognition algorithms engineered by the authors have been introduced in previous publications [20, 24].

Some commercial systems also exist for the recognition of threatening events (especially gunshots). These systems, such as presented by Boomerang [37], ShotSpotter [39] or SENTRI [38], incorporate acoustic event detection and localization to provide information about the location of the shooter. They utilize an array of acoustic pressure sensors as the data source and recurrent neural networks for classification. Such systems are designed to be used in battlefield conditions. They take into consideration two main features of the acoustic event: muzzle blast and shock wave produced by the bullet. Moreover, such systems include several numbers of acoustic sensors, fixed (or mobile) node station and small sensor that can be handled by the soldier. The sensor also include the GPS receivers and wireless communication module. The final result of the position of the shooter can be calculated on the basis of data coming from grid of sensors [9, 10]. Another commercially available example of the practical application of the shooter localization system for military application is the Stand Alone Gunshot Detection Vehicle System [29]. The system also includes acoustic pressure sensors. All these systems were designed and optimized for shooter detection and localization.

In our approach we extended the considered types of sound sources. We also were concentrated on civil application rather than military ones. As it was mentioned before, the systems presented above use the acoustic pressure sensors (microphones). In our approach we use the very small and compact 3D sound intensity probe (Acoustic Vector Sensor—AVS) [6]. This kind of sensors were first applied to acoustic source localization in the air by Raangs et al. in 2002, who used measured sound intensity vector to localize a single monopole source [34]. A more recent development is the application of acoustic vector sensors to the problem of localizing multiple sources in the far field. In 2009, Basten et al. applied the MUSIC method to localize up to two sources using a single acoustic vector sensor [1]. Wind et al. applied the same method to localize up to four sources using two acoustic vector sensors [43, 44].

The authors' experiences with the sound source localization based on the sound intensity methods performing in the time domain or in the frequency domain were presented in details in the previous papers [18–20]. In this paper the authors focus on combining their experience with various algorithms to propose a solution which offers full functionality of: detection, classification and localization the acoustic events in real acoustic conditions. The authors have tested their design in several practical implementations, for example in bank operating room [17]. In the present work we concentrate on preparing the setup for testing our design in various and precisely controlled acoustic conditions. Especially we control three factors: first was the type of background disturbing noise, second—the signal to noise ratio concerning the disturbing noise and considered acoustic events and the final factor was the direction of arrival of radiated acoustic events.

Our engine is meant to be a universal and adaptive solution which can work in low- and high noise conditions, both indoors and outdoors. It is employed in the acoustic monitoring of hazardous events in an audio-visual surveillance system. The information about detected events and their type can be used to inform the operator of the surveillance system of potential threats. In a multimodal application the calculated direction of arrival of the detected acoustic event is used to control the PTZ (Pan-Tilt-Zoom) camera [18, 19]. Thus, the camera is automatically directed toward the localized sound source. The system is designed to operate in real time, both in indoor and outdoor conditions. Therefore, the changing acoustic background is a significant problem. Consequently, the impact of added noise on the performance of the algorithms employed needs to be examined in order for our research to progress.

Most of the published works, known to the authors of this paper, are based on experiments with a database of recorded sounds. For example, in the research by Krijnders et al. [21] a database of self-recorded samples is used, whereas Valenzise et al. utilize events from available sound libraries [42]. Some researchers address the problem of real-world event detection [30, 46]. In such a case the noise added to the signals has to be considered. The most common approach is to mix sounds with recordings of noise digitally, as it was carried out by Mesaros et al. [28] or Lojka et al. [22] In our opinion, it is a different case when the noise is mixed with the signal acoustically (in the acoustic field, thus not being added to the electronic representation of the signal). Therefore in our work we designed an experiment which enables the evaluation of such a case. Our experiments also allow for a more precise estimation of the Signal-to-Noise Ratio (SNR) than it was achieved, to our knowledge, in any of the related work presented in the literature.

The paper is organized as follows. In Section 2 we present our algorithms and methods for detection, classification and localization of acoustic events. In Section 3 we introduce the setup of the experiment and specify the conditions under which the measurements were performed and the equipment used. In Section 4 we discuss the measurement results, leading to the conclusions presented in Section 5.

2 Methods

Commonly, the term Acoustic Event Detection (AED) refers to the whole process of the identification of acoustic events. We divide this process into three phases: detection, classification and localization. The general concept of sound recognition and localization system is presented in Fig. 1. The purpose of detection is to discern between the foreground events and the acoustic background, without determining whether an event is threatening or not. Some researchers use foreground/background or silence/non-silence classifiers to achieve this task [40, 42]. We employ dedicated detection algorithms which do not require training and are adaptive to changing

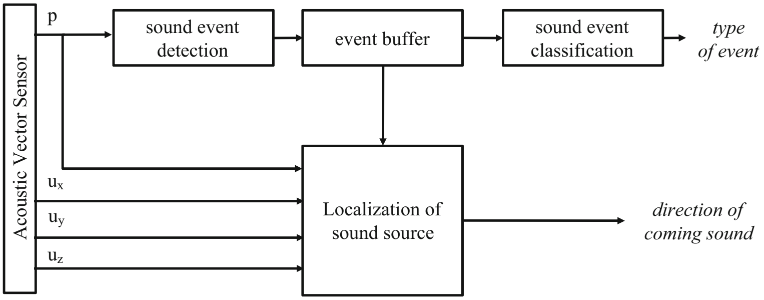

Fig. 1 Concept diagram of a sound detection, classification and localization system

conditions. The detection of a foreground event enables classification and localization, after buffering the samples of the detected event. This architecture enables maintaining a low rate of false alerts, owing to the robust detection algorithms, which we explain in more detail in the following subsections. The classification task is the proper assignment of the detected events to one of the predefined classes. In addition, the localization of the acoustic event is computed by analyzing the multichannel output of the Acoustic Vector Sensor (AVS). The employment of AVS and incorporation of the localization procedure in the acoustic surveillance system provide an addition to the state of the art in sound recognition technology. Stemming from acoustic principles, beamforming arrays have limitations in low frequencies and require line (or plane) symmetry. Data from all measurement points have to be collected and processed in order to obtain the correct results. The acoustic vector sensor approach is broadband, works in 3D acoustical space, and has good mathematical robustness [7]. The ability of a single AVS to rapidly determine the bearing of a wideband acoustic source is essential for numerous passive monitoring systems. The algorithms operate on acoustic data, sampled at the rate of 48,000 samples per second with a bit resolution equal to 32 bits per sample.

2.1 Sound event detection

The conceptual diagram of the sound event detection algorithm is presented in Fig. 2. Initially the detector is set to learning mode. After the learning phase is completed, the detection parameter is compared to the threshold value. This operation yields a decision: "detection" or "no detection". The threshold (or acoustic background profile) is constantly updated to adapt to changing conditions.

We assume that a distinct acoustic event has to manifest itself by a dissimilarity of its features from the features of the acoustic background. The choice of features to be taken into consideration depends on the type of event we intend to detect. This yields four detection techniques:

- based on the short-time level of the signal – applied to detecting sudden, loud impulsive sounds – named Impulse Detector;
- based on the harmonicity of the signal – applied to detecting speech and scream-like sounds – named Speech Detector;
- based on changes in the signal features over time – applied to detecting sudden narrow-band changes in the analyzed signal – named: Variance Detector;
- based on the overall dissimilarity in the spectra of the event and background – applied to detecting any abnormal sounds – named Histogram Detector (since it employs a histogram of sound level in 1/3-octave frequency bands to model the spectrum of the acoustic background).

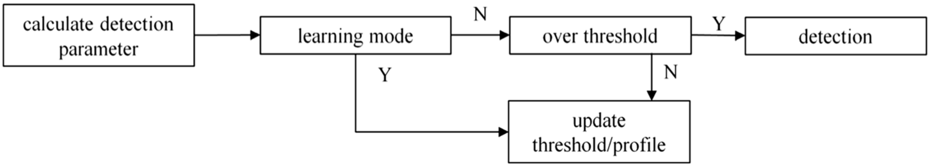

Fig. 2 Conceptual diagram of sound event detection

In general, all detectors rely on comparing the detection parameter P with the threshold T. Hence, the detection function D can be defined as follows:

$$D(i) = \begin{cases} 1 & P(i) > T(i) \\ 0 & P(i) \leq T(i) \end{cases} \tag{1}$$

where i is the index of the current frame. The threshold T is automatically updated to the changes in the acoustic background by exponential averaging according to the formula:

$$\begin{aligned} &T(0) = P(0) + m \\ &T(i > 0) = (1-\alpha)\cdot T(i-1) + \alpha\cdot(P(i) + m) \end{aligned} \tag{2}$$

where m is the margin added to the value of the detection parameter, which serves as a *sensitivity parameter* of the detector. If the detection parameter changes exponentially, m can be a multiplier. The constant α is related to the detector's adaptation time. The adaptation time T_{adapt} is the period after which the previous values of the detection parameter are no longer important. It is related to the constant α according to Eq. 3:

$$T_{adapt}[s] = \frac{N}{SR\cdot\alpha} \tag{3}$$

where N is the number of samples in the frame and SR is the sampling rate. The different detection algorithms employed differ in the definition of the detection parameter and the frame sizes employed. The *Impulse Detector* is based on the level of the signal in short frames (10 ms) calculated as:

$$L = 20\cdot\log\left(\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x[n]\cdot L_{norm})^2}\right) \tag{4}$$

where $x[n]$ are the signal samples and L_{norm} is the normalization factor which equals the level of the maximum sample value measured with a calibration device. *Speech Detector* is based on the *Peak-Valley-Difference* (PVD) parameter. The feature used is a modification of the parameter proposed by Yoo and Yook [45] and often used in *Voice Activity Detection* (VAD) algorithms. The PVD is calculated as follows:

$$PVD = \frac{\sum_{k=1}^{N/2} X(k)\cdot P(k)}{\sum_{k=1}^{N/2} P(k)} - \frac{\sum_{k=1}^{N/2} X(k)\cdot(1-P(k))}{\sum_{k=1}^{N/2}(1-P(k))} \tag{5}$$

where $X(k)$ is the power spectrum of the signal's frame, N=4096 is the length of the Fourier Transform (equal to the length of the detector's frame) and $P(k)$ is a function which equals 1 if k is the position of the spectral peak, 0 otherwise. For typical signals, the spacing of spectral peaks depends on the fundamental frequency of the signal. Since this detection parameter is dedicated to the detection of vocal activity (e.g., screams) the PVD is calculated iteratively over a range of assumed peak spacing corresponding to the frequency range of human voice. Subsequently the maximum value is taken into consideration.

In turn, the *Variance Detector* is based on the variance of signal's features calculated over time. The feature variance vector $Var_f = [\, V_{f1} \quad V_{f2} \quad \dots \quad V_{fN} \,]$ comprises the variances of a total of N signal features. For the n-th feature f_n the feature variance is calculated according to the formula:

$$V_{fn} = \frac{1}{I} \sum_{i=1}^{I} \left(f_n(i) - \overline{f_n} \right)^2 \tag{6}$$

where I is the number of frames used for calculating the variance, i.e., the length of the *variance buffer.* V_{fn} is then used as a detection parameter. The decision is made independently for each feature and the final decision is a logical sum of each feature's detection result. The variance detector is suitable for detecting narrow-band events, since it reacts to changes in single features, some of which reflect the narrow-band characteristics of the signal.

The final detection algorithm is based on a histogram model of acoustic background. The spectral magnitudes are calculated in 1/3-octave bands to model the noise background. 30 bands are used, and for every band a histogram of sound levels is constructed. The detection parameter d_{hist} is then calculated as a dissimilarity measure between the spectrum of the current frame X and the background model:

$$d_{hist}(X) = \sum_{k=1}^{30} h_k(X_k) \tag{7}$$

where $h_k(X_k)$ is the value of the histogram of spectral magnitude in the k-th band. The signals whose spectrum matches the noise profile yield high values of d_{hist}. The histogram-based detection algorithm is designed to deal best with wide-band acoustic events, whose spectral dissimilarity from the acoustic background is the greatest. The algorithm is similar to the GMM detection algorithm, only does not assume Gaussian distribution of sound levels.

2.2 Feature extraction

The elements of the feature vector were chosen on the basis of statistical analysis. Firstly, a large vector of 124 features is extracted from the training set. This large feature vector comprises MPEG-7 descriptors [14], spectral shape and temporal features [32], as well as other parameters related to the energy of the signal, which were developed within a prior work [47]. Secondly, a feature selection technique suited to SVM classification is employed to rank the features. This task is performed using the WEKA data mining tool [27]. We choosed this attribute selection algorithm by briefly comparing it to the other selection methods available in WEKA, namely χ^2 and information gain. In the literature there is a multitude of methods for feature selection, i.a. those introduced by Kiktova [13]. The top 50 features in the ranking are chosen to form the final feature vector. The length of the feature vector was chosen by minimizing the error in the cross-validation check. The composition of the feature vector is presented in Table 1.

Table 1 Elements of the feature vector

Symbol	Feature	Number of features
MPEG-7 spectral features		
ASC	Audio spectrum centroid	1
ASS	Audio spectrum spread	1
ASE	Audio spectrum envelope	20
SFM	Spectral flatness measure	17
Temporal features		
ZCD	Zero crossing density	2
TC	Temporal centroid	1
Other features		
SE	Spectral energy	4
CEP	Cepstral energy	1
PVD	Peak-valley difference	1
TR	Transient features	2

2.2.1 Spectral features

The spectral features are derived from the power spectrum of the signal. The power spectral density function was estimated by employing Welch's method. We will refer to the power spectrum as *P*(*k*), where k denotes the DFT index or *P*(*f*) where f indicates the frequency. The frequency is in this case discrete and relates to the spectral bins according to the formula $f=k\cdot f_s/N$, where f_s equals the sample rate and *N* equals the number of DFT points. The Audio Spectrum Centroid feature is calculated as 1st order normalized spectral moment according to Eq. 8.

$$ASC = \frac{\sum_f P(f)\cdot f}{\sum_f P(f)} \tag{8}$$

The Audio Spectrum Spread Parameter equals the 2nd order normalized central spectral moment and is calculated according to Eq. 8:

$$ASS = \frac{\sum_f P(f)\cdot(f-ASE)^2}{\sum_f P(f)} \tag{9}$$

The Audio Spectrum Envelope group of features expresses the signal's energy in 1/3-octave bands relative to the total energy. Provided that the limits of the 1/3-octave band equal k_1 and k_2, the ASE feature in *m*-th band can be extracted according to Eq. (9):

$$ASE_m = \frac{\sum_{k_1}^{k_2} P(k)}{\sum_k P(k)} \tag{10}$$

A total of 24 1/3-octave bands are taken into consideration. A number of 20 ASE coefficients are then chosen to be included in the feature vector. The next descriptor, the

Spectral Flatness Measure, contains the information about the shape of the power spectrum. The SFM features yield values close to 1 when the signal is noise-like and close to 0 when the signal has some strong harmonic components. Similarly to the ASE calculation, the parameter is extracted in 1/3-octave bands. Equation 10 presents a formula for calculating the spectral flatness of the m-th band, which is employed in this work. Out of the 24 1/3-octave bands 17 SFM coefficients are included in the feature vector.

$$SFM_m = \frac{\prod_{k_1}^{k_2} P(k)^{\frac{1}{k_2-k_1}}}{\frac{1}{k_2-k_1}\sum_{k_1}^{k_2} P(k)} \tag{11}$$

Another group of features comprises spectral energy parameters, which are defined as a ratio of energy in two frequency bands. The limits of the frequency bands are established within a previous work and they match the representative regions in the spectra of different types of acoustic event [47]. Assuming that the first frequency band spans from f_1 to f_2 and the second frequency band spans from f_3 to f_4, the spectral energy feature is calculated according to Eq. 11.

$$SE = \frac{\sum_{f_1}^{f_2} P(f)}{\sum_{f3}^{f_4} P(f)} \tag{12}$$

In the experiments related to this work, 4 spectral energy parameters are included in the feature vector. The respective frequency bands are shown in Table 2.

The last of the spectral parameters is the Peak-Valley Difference (PVD). The PVD relates to the distance between peaks and troughs in the power spectrum. The formula for the calculation of this feature has already been presented (in Eq. 5).

2.2.2 Temporal and cepstral features

The temporal features are extracted from the time-domain representation of the signal, which is referred to as $x[n]$, where n is the sample index. Zero crossing density is a useful temporal feature which reflects the noisiness of the signal. The ZCD parameter is calculated according to the formula:

$$ZCD = \frac{1}{2N}\sum_{n=2}^{N} |sign(x[n]) - sign(x[n-1])| \tag{13}$$

Table 2 Band limits for spectral energy features

Feature	f_1 [Hz]	f_2 [Hz]	f_3 [Hz]	f_4 [Hz]
SE1	1300	1700	0	24,000
SE2	4000	7000	0	24,000
SE3	7000	12,000	0	24,000
SE4	100	500	7000	12,000

where N denotes the total number of samples in the signal. The next temporal feature—temporal centroid of the signal – is calculated according to Eq. 13.

$$TC = \frac{1}{N}\sum_{n=1}^{N} n \cdot x[n] \tag{14}$$

The next feature group is the Cepstral Energy features. The features are derived from the power cepstrum, which is obtained as (Eq. 14)

$$C(n) = F\{\log|F(x)|\} \tag{15}$$

where F denotes the Fourier transform. The cepstral energy features are then calculated by comparing the energy of the part of the cepstrum (i.e.,1/4 of the quefrency axis) with the total energy:

$$CEP_m = \sqrt{\frac{\sum_{n_1}^{n_2} C^2(n)}{\sum_{n} C^2(n)}} \tag{16}$$

where n_1 and n_2 denote the limits of the m^{th} band. The features are extracted from 4 bands and 1 parameter ($0<n\leq 255$) is chosen to be included in the feature vector. The last of the temporal features are transient-related parameters. Two parameters are defined: transient length and transient rate. Both are derived from the first order difference of the signal (referred to as $d[n]$). To detect the transient, the maximum of the first order difference is sought (d_{max}). Then the end of the transient is located by detecting the point at which $d[n]$ falls below the threshold equal to $0.05 \cdot d_{max}$. Once the starting point (n_{tr_start}) and the end point of the transient (n_{tr_stop}) are found, the transient length feature is calculated by subtracting these two values.

$$tr\_length = n_{tr\_start} - n_{tr\_stop} \tag{17}$$

The transient rate feature is defined as the energy ratio of the fragment containing the transient start point and the transient end point (Eq. 17):

$$tr\_rate = 10\log\frac{E(n_{tr\_start})}{E(n_{tr\_stop})} \tag{18}$$

where *E(n)* is the energy in the frame located around index *n*. A 25 ms analysis window was employed for the energy calculation.

2.3 Classification

The system recognizes 4 classes of threatening events and 1 non-threatening event class. In the training set we collcted: 44 explosions, 193 sounds of breaking glass, 676 gunshots, 65 screams and 239 other sounds. The event samples were recorded with the Bruel & Kjaer PULSE system type 7540 in natural conditions, although with a low level of additive noise. Hence, they will hereafter be recognized as *clean* sound events. The files are stored in 48,000 Hz 32-bit floating point WAVE files (the actual bit depth equals 24).

The classification algorithm is based on the Support Vector Machine (SVM) classifier. The principles of SVM and its application to numerous fields have been studied in the literature, namely to text classification [11], face detection or acoustic event detection [35, 40]. It was proven in previous work that the Support Vector Machine can be an efficient tool for the classification of signals in an audio-based surveillance system, as it robustly discerns threatening from non-threatening events [25]. The difficulty pertaining to the employment of SVMs for acoustic event recognition is that SVM, being a non-recurrent structure, is fed a representation of the acoustic event in the form of a static feature vector. Since the length of environmental audio events can vary from less than 1 second to even more than 10 seconds, a correct approach is to divide the signal into frames, classify each frame separately and subsequently make the decision. Such an approach was proposed by Temko and Nadeau [40]. In our work a frame of 200 ms in length is used and the overlap factor equals 50 %. The SVM model employed enables multi-class classification via the 1-vs-all technique with the use of LIBSVM library written in C++ [3]. The model was trained using the Sequential Minimal Optimization method [33]. A polynomial kernel function was used. The output of the classifier, representing the certainty of the classified event's membership in respective classes, can be understood as a probability estimate:

$$P_i(x_n) = SVM\{F(x_n), i\} \tag{19}$$

where P_i is the probability of the analyzed frame x_n belonging to class *i*. F denotes the feature calculation function. The final decision points to the class that maximizes the classifier's output. Moreover, a predetermined probability threshold for each class has to be exceeded. The probability threshold enables the control of false positive and false negative rate. In decision systems theory this problem is known as *detection error tradeoff* (DET). In Fig. 3 the DET curves obtained for the signals from the training set are presented. The optimum threshold is the one that minimizes the loss, i.e., it provides equal error rate (EER). When the rate of false positive results equals the false negative rate, the system operates in minimum-cost configuration. On the plot, it is the point in which the solid line crosses the dashed line. The approximate EERs obtained are: 0.13 for explosion, 0.05 for broken glass, 0.015 for gunshot and 0.017 for scream. The class probability thresholds which yield those EERs are considered optimum being equal to: 0.1 for explosion, 0.45 for broken glass and 0.75 for both gunshot and scream (Fig. 3).

The training procedure comprises the calculation of features from all signals in the event database as well as solving the Support Vector problem, which is performed by employing the Sequential Minimal Optimization algorithm (SMO) [7]. Finally, a cross-validation check is performed, with 3 folds, to assess the assumed model and to evaluate the training of the classifier. The results of the cross-validation check are presented in the form of a confusion matrix in Table 3. The Support Vector classifier yields a very high accuracy on clean signals from the training set.

Even though in this work only 4 selected types of acoustic events are considered, our methods are not constrained to those sound types, only. The employed methodology can be easily adapted to detect and to classify other types of events. For example, in related authors' work the events occurring in a bank operating hall were detected [17].

2.4 Sound source localization

The single acoustic vector sensor measures the acoustic particle velocity instead of the acoustic pressure, which is measured by conventional microphones [12]. It measures the velocity of air particles across two tiny resistive strips of platinum that are heated to approx. 200 °C. It operates in a flow range of 10 nm/s up to ca. 1 m/s. A first order approximation shows no cooling of the sensors, however particle velocity causes the temperature distribution of both wires to change. The total temperature distribution causes both wires to differ in temperature.

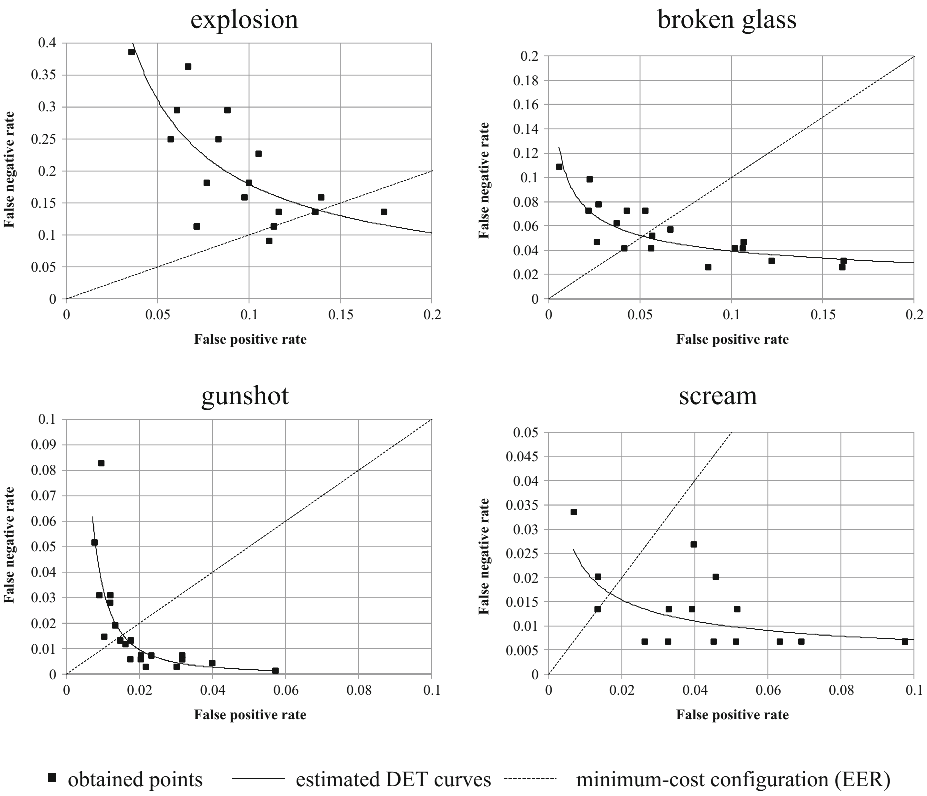

Fig. 3 DET plots for classifying the acoustic events and leading to finding the equal error rate

Because it is a linear system, the total temperature distribution is simply the sum of the temperature distributions of the two single wires. Due to convective heat transfer, the upstream sensor is heated less by the downstream sensor and vice versa. Due to this operation principle, the sensor can distinguish between positive and negative velocity directions and it is much more sensitive than a single hot wire anemometer, and since it measures the temperature difference the sensitivity is (almost) not temperature sensitive [41].

Each particle velocity sensor is sensitive only in one direction, so three orthogonally placed particle velocity sensors have to be used. In combination with a pressure microphone, the sound

Table 3 Cross-validation check of the training procedure

Class:	Classified as:						
	Explosion	Broken glass	Gunshot	Scream	Other	Precision	Recall
Explosion	43	0	1	0	0	1.00	0.98
Broken glass	0	189	2	0	2	0.98	0.98
Gunshot	0	1	675	0	0	0.99	1.00
Scream	0	0	2	63	0	0.94	0.97
Other	0	3	3	4	229	0.99	0.96
Correct classifications / all events (accuracy)						1199/1217	(98.52 %)

field in a single point is fully characterized and the acoustic intensity vector, which is the product of pressure and particle velocity, can also be determined [2]. This intensity vector indicates the acoustic energy flow. With a compact probe, the full three-dimensional sound intensity vector can be determined within the full audible frequency range of 20 Hz up to 20 kHz.

The intensity in a certain direction is the product of sound pressure (scalar) $p(t)$ and the particle velocity (vector) component in that direction $u(t)$. The time averaged intensity I in a single direction is given by Eq. 20 [15].

$$I = \frac{1}{T}\int_{T} p(t)u(t)dt \tag{20}$$

In the algorithm presented the time average T was equal to 4096 samples (sampling rate was equal to 48,000 S/s). It means that the direction of the sound source was updated more than 10 times per second. It is important to emphasize that using the 3D AVS presented, the particular sound intensity components can be obtained solely based on Eq. 19. The sound intensity vector in three dimensions is composed of the acoustic intensities in the three orthogonal directions (*x,y,z*) and is given in Eq. 19 [15] .

$$\overrightarrow{I} = I_x \overrightarrow{e}_x + I_y \overrightarrow{e}_y + I_z \overrightarrow{e}_z \tag{21}$$

The authors' experience with the sound source localization based on sound intensity methods performed in the time domain or in the frequency domain is presented in their previous papers [15, 16], whereas the algorithm for acoustic events localization applied during this research operates in the time domain. Its functionality was adapted to work with detection and classification algorithms. The direction of arrival values are determined on the basis of acoustical data available in event buffer (see Fig. 1, Sec. 2). The angle of the incoming sound in reference to the acoustic vector sensor position is the main information about the sound source position. For a proper determination of the sound source position, the proper buffering of the acoustic data and a precise detection of the acoustic event are needed.

Such a process enables the proper selection of the part of the sound stream which includes the data generated by the sound source. Only samples buffered for detected acoustical event are taken into account during the computing of the sound intensity components. Acoustic events used in the experiment executed had a different length. For that reason the buffered sound samples of the detected acoustic event were additionally divided into frames of 4096 samples. For each frame the sound intensity components and angle of the incoming sound were calculated. The functionality and some additional improvements of the localization of sound sources algorithm for the application in real acoustic conditions can be found in related works [15, 16].

3 Experiment

In the experiment we make an attempt to evaluate the efficiency of detection, classification and localization of acoustic events in relation to the type and level of noise accompanying the event. These are: traffic noise, railway noise, cocktail-party noise and typical noise inside buildings. The key parameter is the Signal-To-Noise Ratio (SNR). We decide to perform the experiments in laboratory conditions, in an anechoic chamber. This environment, however far from being realistic, gives us the possibility to precisely control the conditions and to measure the levels of sound events and noise, which is substantial in this experiment. It also eliminates the room reflections, thus simulating an outdoor environment. The drawback of this approach is that the signals reproduced by speakers are used, instead of real signals, which has its impact both on the recognition and localization of events. The setup, equipment utilized and the methodology of the conducted experiment are discussed in detail in the following subsections.

3.1 Setup and equipment

The setup of the measurement equipment employed in the experiment is presented in Fig. 4. In an anechoic chamber, 8 REVEAL 601p speakers, an USP probe and a type 4189 measurement microphone by Bruel & Kjaer (B&K) were installed. The USP probe is fixed 1.37 meters above the floor. The measurement microphone is placed 5 mm above the USP probe. In the control room a PC computer with Marc 8 Multichannel audio interface is used to generate the test signals and record the signals from the USP probe. Two SLA-4 type 4-channel amplifiers are employed to power the speakers. In addition, PULSE system type 7540 by B&K is used to record the acoustic signals. The PULSE measuring system is calibrated before the measurements using a type 4231 B&K acoustic calibrator. The angles (α) and distances (d) between the speakers and the USP probe are listed in Table 4. The speakers were placed at 1.2 m height. The angular width of the speakers ($\Delta\alpha$) was also measured. Detailed placement of speakers and real view of the experiment setup are additionally presented in Fig. 5a and b.

3.2 Test signals

Audio events were combined into a test signal consisting of 100 events, randomly placed in time, 20 examples of each of the 5 classes. The average length of each event equals 1 second, and there is a 10 second space between the start and end of adjacent events. The length of the test signals equals 18 min 20 s. Four disturbing signals were prepared, each with a different type of noise:

- traffic noise, recorded in a busy street in Gdansk;
- cocktail-party noise, recorded in a university canteen;
- railway noise, recorded in Gdansk railway station;
- indoor noise, recorded in the main hall of Gdansk University of Technology.

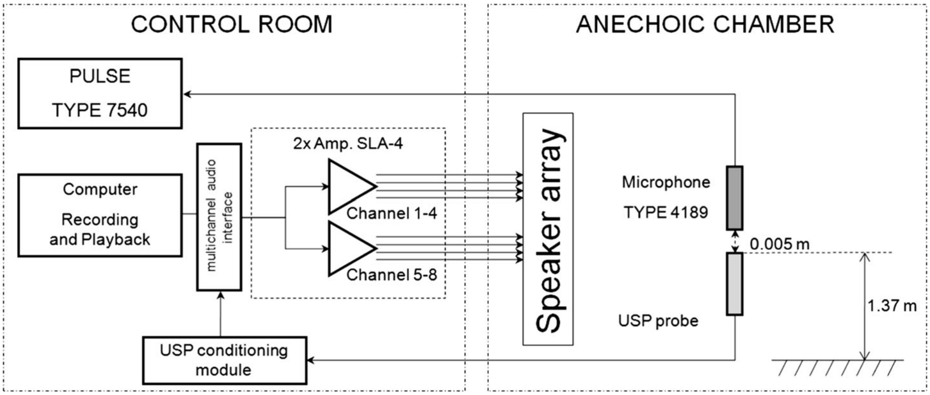

Fig. 4 Experiment setup diagram

Table 4 Angles and distances between the speakers and USP probe/microphone

Speaker no.	Distance (d) [m]	Angle (α) [°]	Angular width ($\Delta\alpha$) [°]
1	1.984	52.1	±3
2	2.182	314.2	±2.5
3	3.476	297.8	±2
4	3.259	255.7	±2
5	1.687	198	±3
6	2.098	143.8	±3
7	2.737	111.6	±2.5
8	3.223	77.2	±2

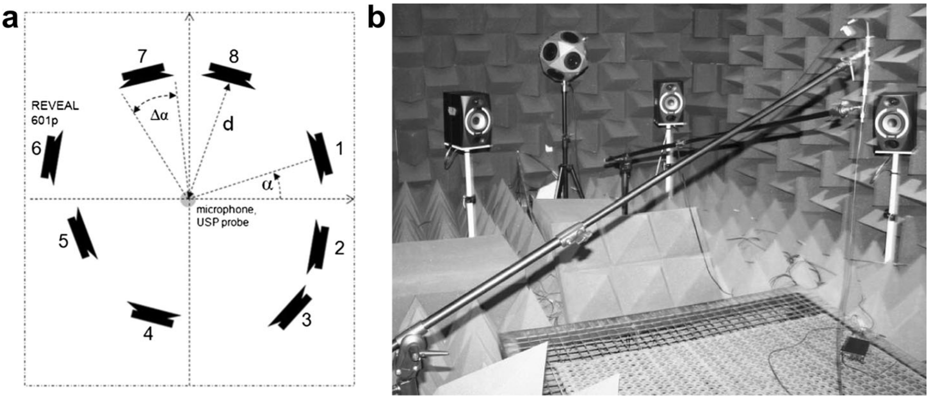

Fig. 5 Placement of speakers in the anechoic chamber (**a**), view of the experiment setup (**b**)

All noise signals were recorded using a B&K PULSE system and were written to 24-bit WAVE files sampled at 48,000 samples per second. Energy normalized spectrums of the particular disturbing sounds are presented in Fig. 6. The differences in energy distribution for used signals are clearly noticeable. The indoor noise has the energy concentrated in the middle part of the spectrum (200 Hz–2000 Hz). The very high level of tonal components for railway noise was produced by the brakes.

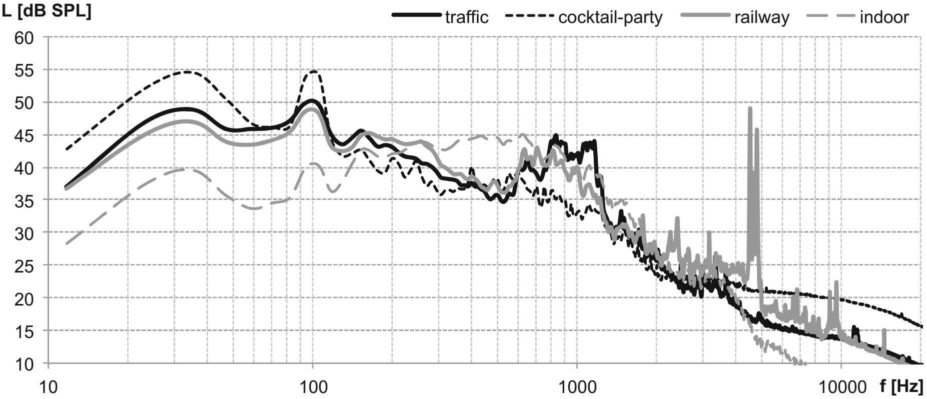

Fig. 6 Energy-normalized spectrum of the particular noise signals used during the experiments

3.3 Methodology

In the test signals the events were randomly assigned to one of four channels: 1,3,5,7 (as defined in Table 5). The order of the events with the numbers of channels they are emitted from and classes they belong to is stored in the Ground Truth (GT) reference list. At the same time, the other channels (2,4,6,8) are used to emit noise. Each noise channel is shifted in time to avoid correlation between channels. The gain of the noise channels is kept constant, while the gain of events is set to one of four values: 0 dB, -10 dB, -20 dB and -30 dB. This yields 16 recordings of events with added noise (4 types of noise x 4 gain levels). In addition, the signals of four types of noise without events and 4 signals of events without noise with different gain levels are recorded. These events are used to measure the SNR. Totally, 23 signals have been gathered (indoor noise at -30 dB gain was later excluded due to too low level). The total length of the recordings equals 7 h 02 min. The summary of the recordings is presented in Table 5.

3.3.1 SNR determination

The exact determination of SNR is a challenging task. In theory SNR is defined as the relation of signal power to noise power. These values are impossible to measure in practical conditions

Table 5 Recordings data

No.	Recording	Events gain	Number of events	Time [hh:mm:ss]
1	Events without noise	0	100	00:18:20
2	Events without noise	−10	100	00:18:20
3	Events without noise	−20	100	00:18:20
4	Events without noise	−30	100	00:18:20
5	Traffic noise only			00:18:20
6	Cocktail-party noise only			00:18:20
7	Railway noise only			00:18:20
8	Indoor noise only			00:18:20
9	Events with traffic noise	0	100	00:18:20
10	Events with traffic noise	−10	100	00:18:20
11	Events with traffic noise	−20	100	00:18:20
12	Events with traffic noise	−30	100	00:18:20
13	Events with cocktail-party noise	0	100	00:18:20
14	Events with cocktail-party noise	−10	100	00:18:20
15	Events with cocktail-party noise	−20	100	00:18:20
16	Events with cocktail-party noise	−30	100	00:18:20
17	Events with railway noise	0	100	00:18:20
18	Events with railway noise	−10	100	00:18:20
19	Events with railway noise	−20	100	00:18:20
20	Events with railway noise	−30	100	00:18:20
21	Events with indoor noise	0	100	00:18:20
22	Events with indoor noise	−10	100	00:18:20
23	Events with indoor noise	−20	100	00:18:20
		Total:	1500	07:02:40

when the noise is always added to the useful signal. Therefore, we propose a methodology of experimentation which allows us to measure the SNR of a sound event. To measure SNR, separate measurements of the sound pressure level were taken, first of events without noise (recordings 1–4 in Table 5), then of noise without events (recordings 5–8 in Table 5). The SNR is calculated by means of the equivalent sound level in the length of the acoustic event (Eq. 20):

$$SNR[dB] = 10\cdot\log\left(\frac{\sum_{k=k_1}^{k_2} s^2[k]}{\sum_{k=k_1}^{k_2} n^2[k]}\right) \tag{22}$$

where $s[k]$ is the signal containing acoustic events and $n[k]$ is the noise signal and $[k_1;k_2]$ is the range of samples in which the acoustic event is present.

The SNR values for particular acoustic events were determined for both signals recorded using the PULSE measuring system and acoustic pressure data recorded by means of the USP probe. SNR data calculated based on signals delivered by the PULSE system give the best information, which can be measured in the open acoustic field. These values were used during the evaluation process of the described sound source localization algorithm. Moreover, these values can be used to determine the sensitivity of the algorithms presented in the dB SPL scale in reference to 20 μPa (for 1 kHz). Additionally, SNR values were determined for signals obtained by means of the USP probe. These values includes the properties of the whole acoustic path, especially the self-noise and distortion, and they reflect the real working condition of the particular algorithms. For further analysis and for the presentation of the results of the sound event detection and classification, the SNR values are divided into the following intervals: {(-∞;-5 dB]; (-5 dB;0 dB]; (0 dB;5 dB]; (5 dB;10 dB]; (10 dB;15 dB]; (15 dB;20 dB]; (20 dB;25 dB]; (25 dB;∞)}.

In Fig. 7, the described methodology of the determination of the SNR values is illustrated. In the first step, the energy of the particular acoustic events was calculated. It is presented in the top chart in Fig. 7. Parts of the signal which include the acoustic events are marked by grey rectangles. In the second step, the energy of the considered background noise level is measured. It is important to emphasize that the background noise levels are determined for synchronous periods of time in relation to particular acoustic events. This means that the noise level that is originating from the acoustic event is not taken into account in the noise level calculations. This is illustrated in the middle chart in Fig. 7. In the bottom chart the particular acoustic events with the background noise considered are plotted. This signal is used during the described analysis. Based on these measurements, we obtain a detailed and precise information about the SNR for each acoustic event.

3.3.2 Detection and classification rates

The experiment recordings are analyzed with the engineered automatic sound event detection and localization algorithms. The measures of detection accuracy are the True Positive (TP), and False Positive (FP) rates. The TP rate equals the number of detected events which match the events in the GT list divided by the total number of events in the GT list. The matching of event is understood as the difference between detection time and GT time of the event being not greater than 1 second. A FP result is considered when an event is detected which is not

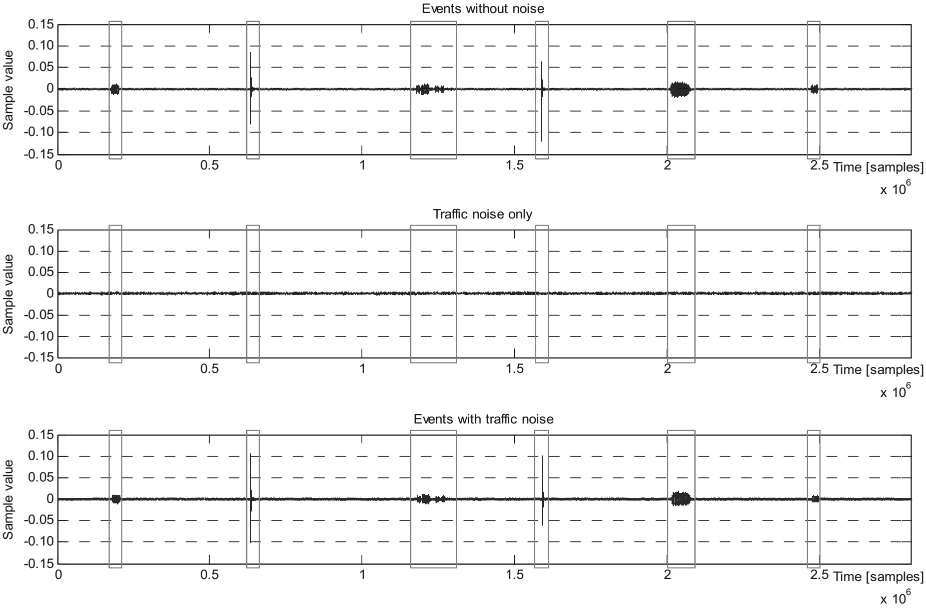

Fig. 7 Illustration of the SNR calculation

listed in the GT reference and is classified as one of the four types of event that are considered alarming (classes 1–4). The assumed measures of classification accuracy are precision and recall rates, which are defined as follows (Eq. 21):

$$\begin{aligned} precision_c &= \frac{number\ of\ correct\ classifications\ in\ class\ c}{number\ of\ all\ events\ assigned\ to\ class\ c} \\ recall_c &= \frac{number\ of\ correct\ classifications\ in\ class\ c}{number\ of\ all\ events\ belonging\ to\ class\ c} \end{aligned} \tag{23}$$

3.3.3 Localization accuracy

The algorithm applied to the determination of the position of the sound source returns the result as a value of the angular direction of arrival. For the determination of the localization accuracy the real positions of the used sound sources in relation to the USP probe are needed. The data are obtained during the preparation of the experiment setup and they are the Ground Truth data of the position of the particular sound source. The reference angle values of the particular loudspeakers are given in Table 4. Taking into consideration the presented assumptions, the sound source localization accuracy (α_{err}) is defined as a difference between the computed direction of arrival (α_{AVS}) angle and the real position of the sound source (α_{GT}). This parameter value is given by Eq. 22:

$$\mathrm{a}_{\mathrm{err}} = \mathrm{a}_{\mathrm{AVS}} – \mathrm{a}_{\mathrm{GT}} \tag{24}$$

The examination of the localization accuracy was performed for all signals and for disturbing conditions described in the methodology section.

4 Results

4.1 Detection results

The results of sound event detection are presented in Fig. 8. The TP rates of each of the detection algorithms vs. SNR are plotted. The combination of all detection algorithms yields high detection rates. The TP rate decreases significantly with the decrease of SNR. The algorithm which yields the highest detection rates in good conditions (SNR >10 dB) is the Impulse Detector. It outperforms the other algorithms, which are more suited to specific types of signal. However, the Impulse Detector is most affected by added noise, since it only reacts to the level of the signal. Other algorithms, namely Speech Detector and Variance Detector, maintain their detection rates at a similar level while SNR decreases. It is a good feature, which allows the detection of events even if they are below the background level (note the TP rate of 0.37 for SNRs smaller than -5 dB). It is also evident that the combination of all detectors performs better than any of them alone, which proves that the engineered detection algorithms react to different features of the signal and are complementary. The Histogram Detector is disappointing, since its initial TP rate is the lowest of all detectors and falls to nearly 0 at 5 dB SNR. The total number of detected events equals 1055 out of 1500 (for all SNRs combined) which yields an average TP rate of 0.7.

In Fig. 9 the TP rate of detection for the different classes of events and types of disturbing noise are presented. On average, the detectors perform best in the presence of cocktail-party noise, compared to other types of disturbing signals. The worst detection rates are achieved in the simulated indoor environment. It can also be observed that some classes of acoustic events are strongly masked by specific types of noise. Gunshots for example have a TP rate of 0.45 in the presence of traffic noise and 0.74 in the presence of railway noise.

The next graph in Fig. 10 shows how different detection algorithms cope with recognizing different types of event. The results are average the TP rates for all values of SNR. The

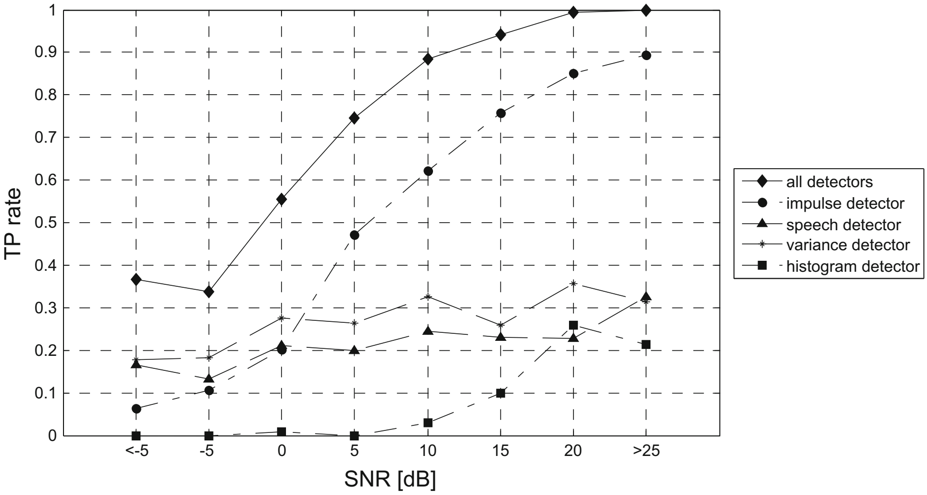

Fig. 8 TP detection rates

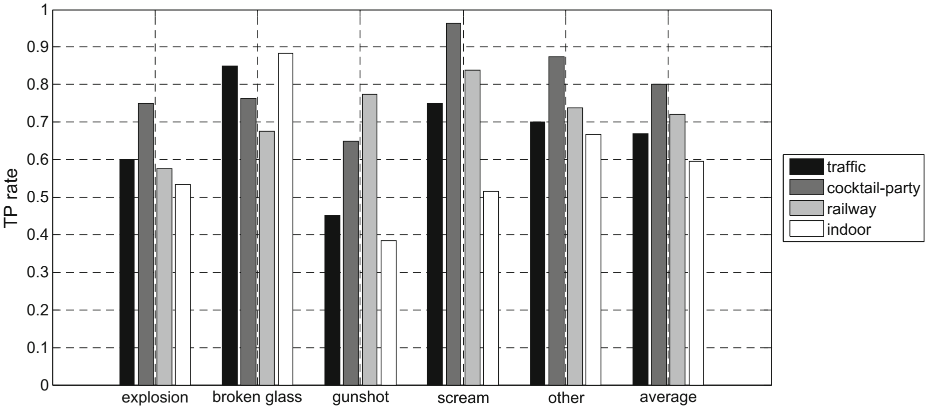

Fig. 9 TP detection rates for different classes of acoustic events and types of noise

presented dependencies once again prove that the developed detection algorithms complement one another and they are suited to recognizing specific types of event. The *Speech Detector* reacts to tonality which is present in screams, while *Variance Detector* reacts to sudden changes in features related to the event of breaking glass. It proves the assumptions made while designing the detectors, which are introduced in Section 2.

A very important aspect, as far as sound event detection is concerned, is false alarms. In our experiment a detection is treated as a FP value when the detected event was not present in the *Ground Truth* reference list and is recognized as one of the classes related to danger (classes 1–4). The number of false alarms produced by each detection algorithm and the classes that are falsely assigned to them are presented in Table 6. The presented FP rates are calculated with respect to the total number of events detected by the specific detector. It can be seen that *Speech Detector* and *Impulse Detector* produce the majority of the false alarms. The fact is understandable, since these algorithms react to the level of the signal and to tonality. Sudden changes in the signal's level and tonal components appear in the acoustic background frequently. The lowest FP rate is achieved by the *Histogram Detector*, however it also yields the lowest TP rate. The *Variance Detector* achieves satisfactory performance, as far as FP rate

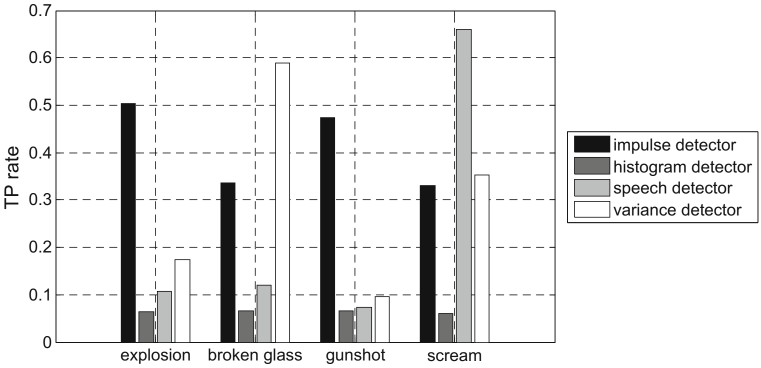

Fig. 10 TP detection rate of events in respect to detection algorithm

Table 6 Number of FP detections

	Impulse detector	Histogram detector	Speech detector	Variance detector	All detectors
Explosion	12	0	1	0	13
Broken glass	26	0	27	8	50
Gunshot	12	0	7	1	20
Scream	3	1	9	1	9
Sum	53	1	44	10	92
FP rate	0.07	0.01	0.12	0.02	0.08

is concerned. It is a good feature, demonstrating the fact that its TP rate is robust against noise. The overall FP rate equals 0.08, which can be regarded as a good performance.

4.2 Classification results

The adopted measures of classification accuracy, i.e., precision and recall rates, were calculated with respect to SNR. The results are presented in Fig. 11.

The general trend observed is that the recall rate descends with the decrease in SNR. It can be seen, as far as explosion and broken glass are concerned, that the precision rate ascends with the decrease in SNR. In very noisy conditions these classes are recognized with greater

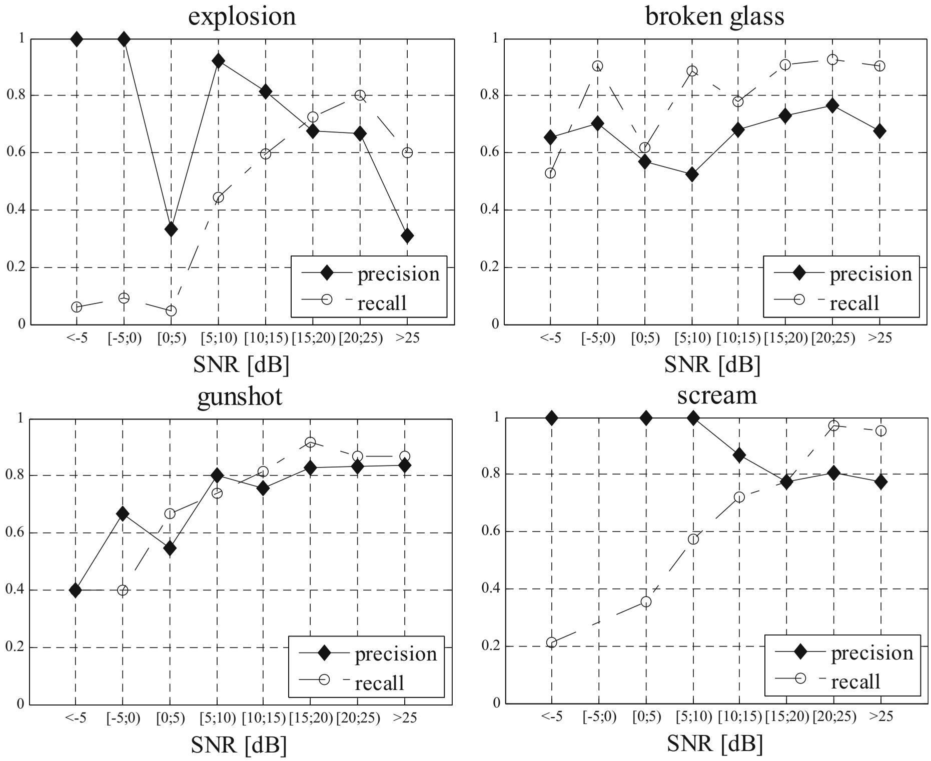

Fig. 11 Precision and recall rates of sound events in relation to SNR

Table 7 Confusion matrix at 20 dB SNR

Class:	Classified as:						
	Explosion	Broken glass	Gunshot	Scream	Other	Precision	Recall
Explosion	24	2	1	0	3	0.67	0.80
Broken glass	0	26	0	0	2	0.76	0.93
Gunshot	1	1	20	0	1	0.83	0.87
Scream	1	0	0	33	0	0.80	0.97
Other	10	5	3	8	12	0.67	0.32
Correct classifications / all events (accuracy)						115/153	(75.16 %)

certainty. The class of event which is least affected by noise is broken glass. The recall rate remains high (ca. 0.8 or more) for SNRs greater than or equal to 5 dB. The low overall recall rate of explosions is caused by the fact that the events were reproduced through loudspeakers, which significantly changes the characteristics of the sound. This aspect is discussed further in the conclusions section. The precision rate for explosions also deserves consideration. It can be noticed that the precision rate achieved for 0 dB SNR does not match the rest of the curve. It is due to the fact that there are very few events classified as explosion for low SNRs. For 0 dB SNR 2 non-threatening events were erroneously classified as explosion, thus dramatically lowering the precision rate (see Table 8). For the lower SNR values such errors were not observed, so the points follow a more predictable pattern.

To examine the event classification more thoroughly, we present more data. In Tables 7 and 8 two confusion matrices are presented—at 20 dB and at 0 dB SNR respectively. It is apparent that when the noise level is high, the threatening events are often confused with other, non-threatening events. The errors between the classes of hazardous events are less frequent. It can also be seen that at 20 dB SNR there are frequent false alarms, especially falsely detected explosions (in 10 cases) and screams (8 cases). In audio surveillance, however, such false alarms should always be verified by the human personnel, therefore such error is not as important as classifying a hazardous event as non-threatening (false rejection).

4.3 Localization results

Two types of analyses of sound source localization results are performed. The first type is related to the presentation of localization accuracy of particular types of acoustic events and

Table 8 Confusion matrix at 0 dB SNR

Class:	Classified as:						
	Explosion	Broken glass	Gunshot	Scream	Other	Precision	Recall
Explosion	1	5	1	0	14	0.33	0.05
Broken glass	0	21	2	0	11	0.57	0.62
Gunshot	0	0	6	0	3	0.55	0.67
Scream	0	6	1	11	13	1.00	0.35
Other	2	5	1	0	16	0.28	0.67
Correct classifications / all events (accuracy)						55/119	(46.22 %)

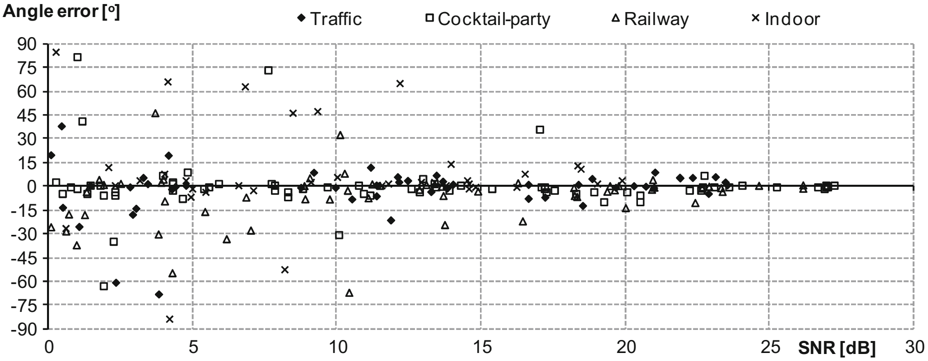

Fig. 12 Localization results for source type: explosion as a function of SNR values for different type of disturbing noise

disturbing noise in relation to the SNR. The second analysis is focused on the determination of localization accuracy in relation to source positions and SNR level.

4.3.1 Localization accuracy in relation to type of acoustic event and disturbing noise

The main aim of this analysis is a direct comparison of how different noise types affect the localization accuracy of the type of sound source considered. Thus prepared graphs are presented in Figs. 12, 13, 14, 15 and 16. On the basis of obtained results we find that the best localization accuracy is observed for non-impulsive sound events like screams and partially broken glass. For this kind of events a proper localization is possible even for SNR at the level of 5 dB. The best localization accuracy is obtained for scream event in the indoor noise. Traffic and railway noise disturbed localization of this events more than cocktail-party and indoor noise. For SNR below 5 dB the localization error increases rapidly.

For impulsive sound events like explosions and gunshots we obtain a proper localization for SNR greater than 15 dB. Below this level the error of localization also grows rapidly. Railway noise has a greater impact on localization of this kind of events than other tested disturbing signals. Gunshot has the best localization accuracy for traffic noise even for SNR

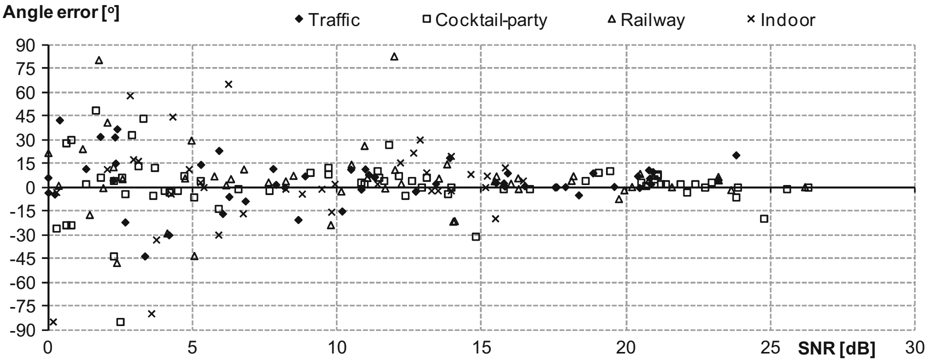

Fig. 13 Localization results for source type: broken glass as a function of SNR values for different type of disturbing noise

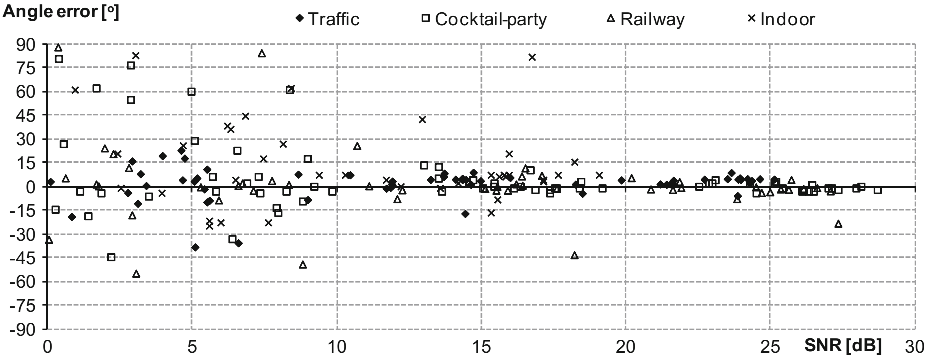

Fig. 14 Localization results for source type: gunshot as a function of SNR values for different type of disturbing noise

about 10 dB. Localization results for source type: explosion as a function of SNR values for different type of disturbing noise.

In Fig. 17 additional results are presented. In this case angular error is calculated for considered types of disturbing noises without division with respect to type of acoustic events.

Localization results calculated for all type of events clearly confirm that railway noise influences the localization accuracy mostly. This is confirmed by the fastest growth of the localization error in relation to SNR level under the same disturbance conditions. In Fig. 18 results for the considered type of acoustic events without distinction between different types of disturbing noise are depicted. The main purpose of this analysis is presentation of relative differences between the localization accuracy for different types of acoustic events. Obtained results confirm that scream is the sound event type which is localized with the best accuracy for SNR up to 5 dB. Other kinds of acoustic events are properly localized when the SNR exceeds 15 dB, ensuring low localization error.

In Fig. 19, the averaged angle localization error as a function of SNR level is presented. The graph is prepared for all recorded acoustic events for every disturbance condition. The events are sorted in order of descending SNR. The angle error curve is averaged with a time constant equal to 15 samples. The whole set contains 1500 events. As indicated above, the significant increase in localization error starts for a SNR level lower than 15 dB.

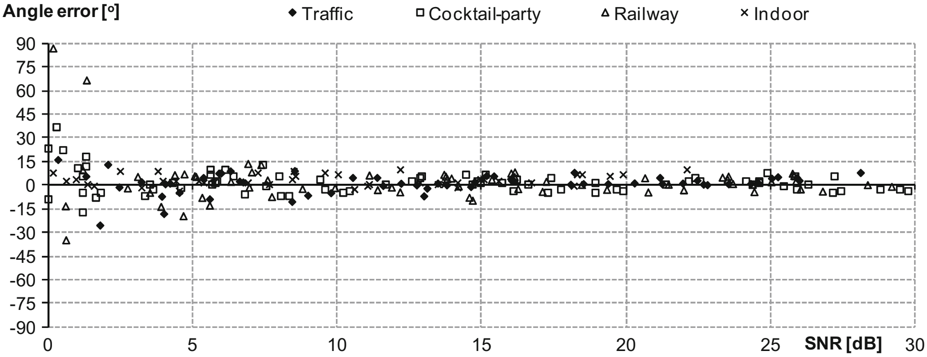

Fig. 15 Localization results for source type: scream as a function of SNR values for different type of disturbing noise

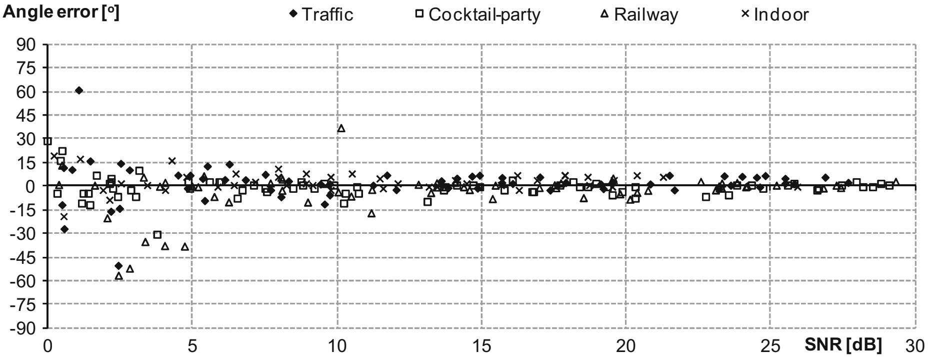

Fig. 16 Localization results for source type: other as a function of SNR values for different type of disturbing noise

4.3.2 Localization accuracy in relation to source position

In this analysis the results obtained are grouped in relation to particular sound sources (i.e., loudspeakers) and presented in Figs. 20 and 21. The true position of the loudspeaker and the localization results are shown in the Cartesian coordinate system. SNR values are indicated by different types of marker and the length of the radius. Distinctions due to the type of event and disturbance noise are not considered in this case. The main purpose of this presentation is the visualization of the distribution of localization error in relation to the SNR level. It is important to emphasize that the loudspeakers employed are not an ideal point source of sound. Every loudspeaker has its own linear dimensions and directivity. These parameters have an influence on the localization results obtained, especially for broadband acoustic events like gunshots, explosions or broken glass. For that reason, in practical situations when the real sound source rapidly emits the high level of acoustic energy, its localization can be even more precisely determined than in the prepared experiments.

Based on localization results obtained, an additional analysis is performed. The values of average error and standard deviation as a function of SNR values are computed. The results are shown in Fig. 22. The mean error is close to 0, but with a decrease in SNR value, the standard deviation increases. For SNR lower than 10 dB the localization decreases rapidly. Figure 23

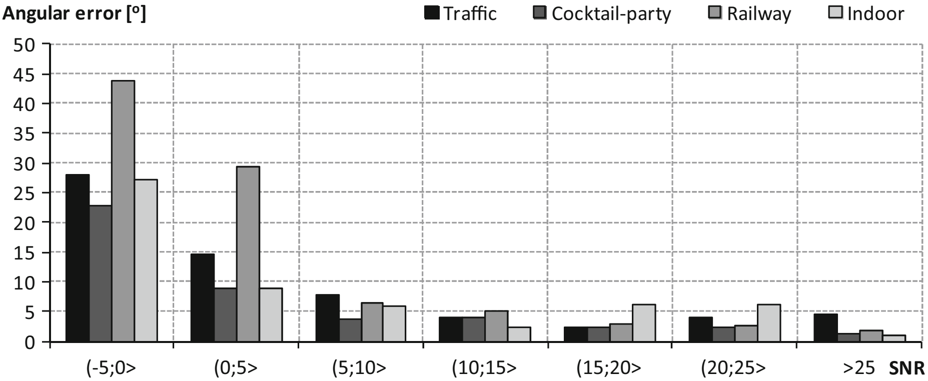

Fig. 17 Localization results (expressed as median values of angular error) for all events plotted as a function of SNR for different types of disturbing noise

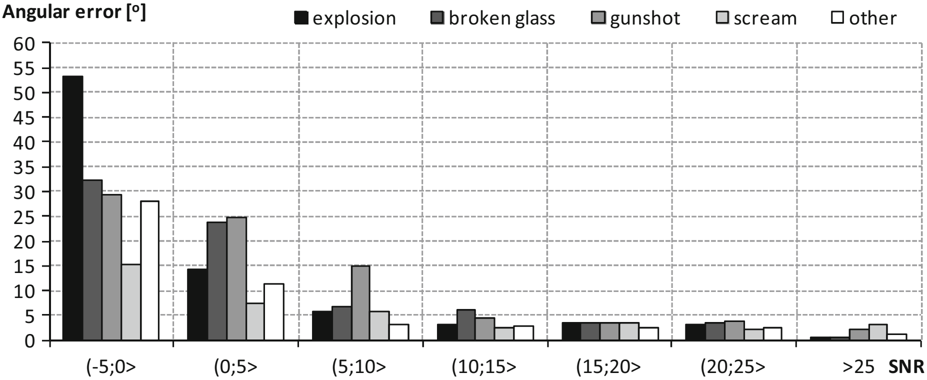

Fig. 18 Localization results (expressed as median values) for all type of noises plotted as a function of SNR values for different types of acoustic events

presents the error values distribution as a function of SNR. The percentage values of correctly localized sound events are also presented. For SNR up to 10 dB almost half the sound events were localized precisely. A decrease in SNR level increases both the probability of inaccurate localization and the error value.

4.4 Real-world experiment

The recognition results need to be discussed with regards to potential real-world applications. The follow-up experiment was organized in which real-world events were emitted in an outdoor environment near a busy street. The results of this experiment have been partially presented in a related conference paper [23]. Real-world examples of glass breaking, scream and shots from the noise gun were used. Explosion sounds were not emitted in the experiment due to technical difficulties in producing them. The microphones were placed in varied distance from the sources of events (2–100 meters), thus yielding similar SNR values to the ones achieved in the anechoic chamber. The results obtained in the real-life experiment follow a very similar trend to the ones achieved in the anechoic chamber. In Table 9 the detection results are presented. The events were detected by a combination of impulse detector and speech detector. The TP detection rates with respect to SNR together with overall TP and FP rates are included in the table. The achieved detection rates vary depending on the event type.

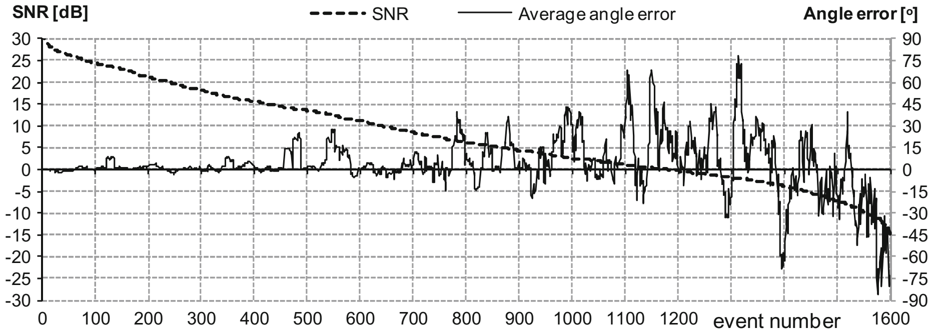

Fig. 19 Localization results for all sound source types as a function of SNR values for indoor noise

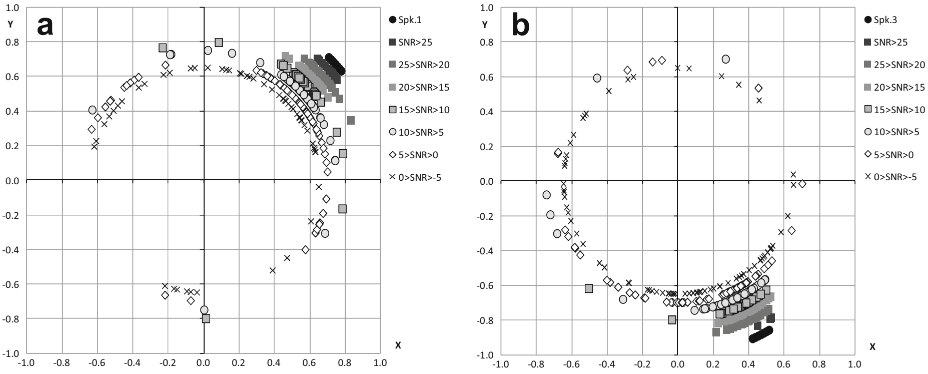

Fig. 20 Sound event detection and localization results: sound events presented from speaker 1 (plot A) and 3 (plot B). Different *shaded dots* indicate the estimated positions for particular SNR values. The *black dots* (for the greatest radius) indicate the true position of the sound source

For the broken glass case a low TP rate is achieved for SNRs smaller than 10 dB. However, the gunshot sounds are detected with a satisfying accuracy even for small SNRs.

Next, in Fig. 24 the precision and recall rates are shown for the considered classes of acoustic events. As it can be seen, the correctly detected events are considered. The obtained plots are similar to the ones shown in Fig. 11. For a more detailed examination of the recognition results a confusion matrix is shown in Table 10. The table aggregates results for all SNR levels. It can be noted that the recall and precision rates are sufficient for identifying hazardous acoustic events in real-world conditions.

Finally, the recall and precision rates achieved in real conditions are directly compared to the ones obtained in the anechoic chamber. In case of real conditions, the SNR was from the range (0;10 dB] and in simulated conditions the SNR falls between 0 and 5 dB. The results are shown in Table 11. It can be observed that the recall and precision rates in real conditions are very close to the ones obtained in the anechoic chamber. In fact, the results are even slightly better in the real-world conditions. This finding can be explained by the fact that in the anechoic chamber the events were reproduced through loudspeakers. In the light of the outcome of the follow-up experiment we can expect that the results discussed in this paper

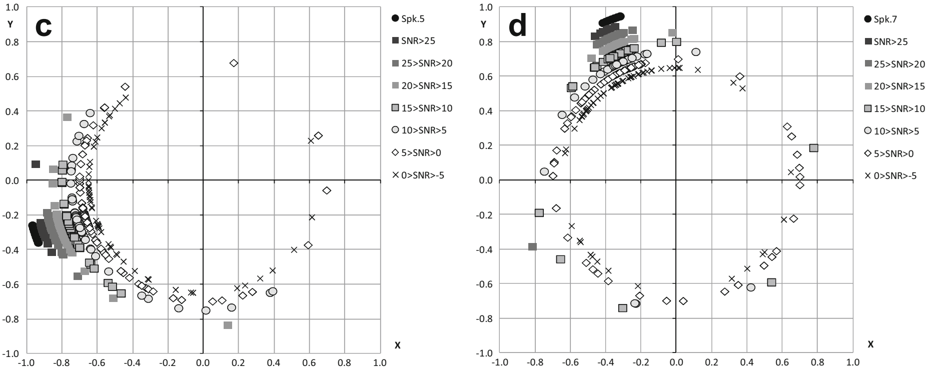

Fig. 21 Sound event detection and localization results, sound events presented from speaker 5 (plot C) and 7 (plot D). Different *shaded dots* indicate the estimated positions for particular SNR values. The *black dots* (for the greatest radius) indicate the real position of the sound source

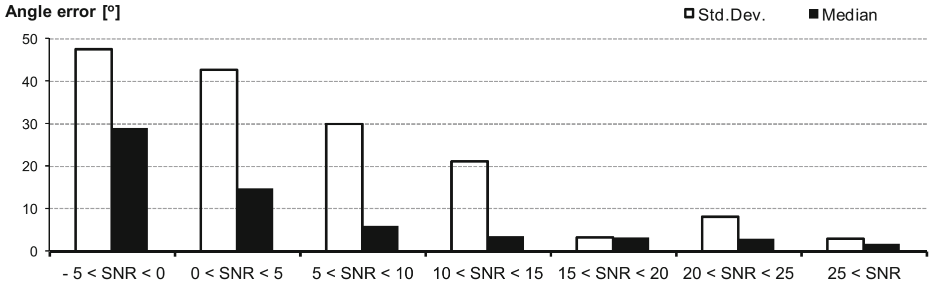

Fig. 22 Average angle error and standard deviation calculated and presented as a function of SNR value

will translate to the real-world cases. It also proves the usefulness of the experiments carried out in the anechoic chamber. The anechoic chamber provides a good simulation of the outdoor conditions, due to very low level of reflections. If the experiment was carried out in a reverberant room, the room acoustics would influence the recognition results and thus the evaluation would not make a universal reference.

5 Conclusions

Methods for automatic detection, classification and localization of selected acoustic events related to security threats have been presented. The algorithms were tested in the presence of noise of different types and intensity. The relations between SNR and the algorithms' performance were examined. The analysis of the results shows that some conditions of the experiment may impair the performance of the methods employed. The most significant limitation is that the acoustic events were played through loudspeakers, whereas the characteristics of sound which is reproduced by speakers (especially dynamic and spectral features) may differ from those of real sounds. This yields a relatively low recall rate for gunshots and explosions. These types of event are practically impossible to be reproduced through speakers with enough fidelity with respect to preserving the dynamics and spectral content of the sound.

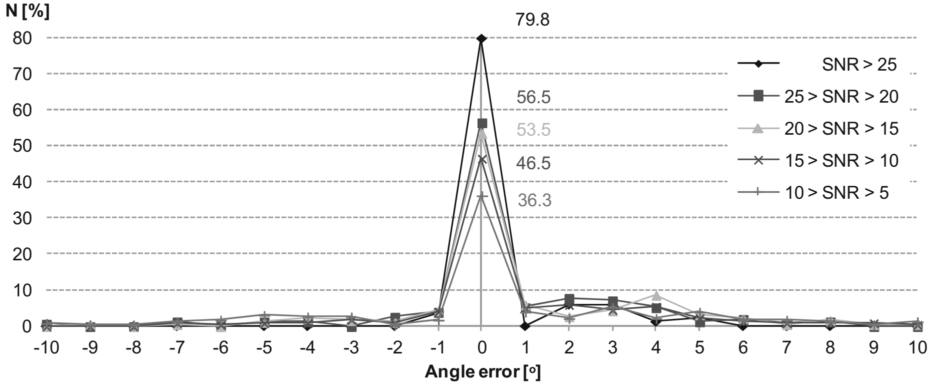

Fig. 23 Error value distribution as a function of SNR value. The percentage values of correctly localized sound events are also presented

Table 9 Detection results in real-world conditions

SNR:	<0	[0;10)	[10;20)	> 20	Overall TP	Overall FP
Broken glass	0.118	0.324	0.932	0.947	0.537	0.225
Gunshot	1	0.98	0.947	1	0.992	0.095
Scream	0.2	0.446	1	1	0.666	0.063
All events	0.708	0.547	0.974	0.996	0.768	0.109

Therefore the training samples, providing recordings of real events, in some cases do not match the signals analyzed within this experiment in the space of acoustic features. The effect is that gunshots and explosions are either confused with non-threatening events, or confused with each other.

The values of SNR in this experiment are realistic, i.e., such SNRs are encountered in environmental conditions. It appears that the precision and recall rates achieved in the cross-validation check performed on the training set are very difficult to achieve in the experiment. The possible reasons for such degraded performance are:

- insufficient noise robustness of features, whose values change significantly when noise is added; evaluation of noise robustness of features should be performed to assess this phenomenon;
- low noise robustness of the classification algorithm (possibly overfitted to clean signals); the classifier's performance should be compared with other structures;
- coincidence of the important spectral components of noise with the components of the events which are substantial for recognizing them (low recall rate of screams in the presence of cocktail-party noise);
- conditions of this experiment, namely reproducing the events through loudspeakers.

These aspects should be examined in future research on the subject in order to improve the noise robustness of the recognition algorithms employed.

The recognition engine was also evaluated in real-world conditions. The performance achieved in the real-world setup is comparable to the results of the laboratory evaluation. It proves that the anechoic chamber makes a good way to simulate conditions of the acoustic environment. Hence, in the light of the achieved results it is to conclude that the results of this work will translate to the real-world case.

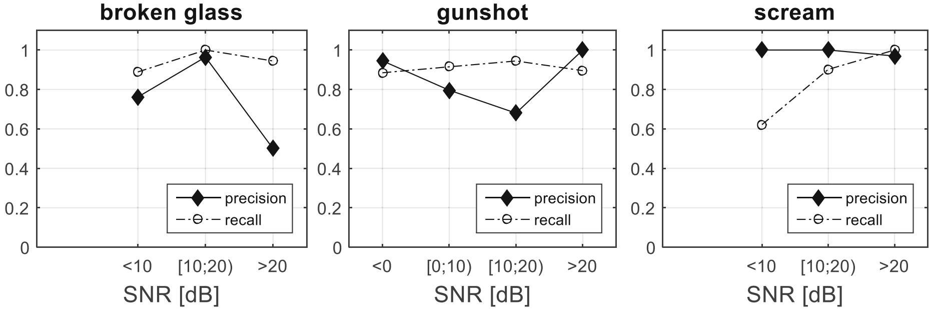

Fig. 24 Precision and recall measures of event classification in real-world conditions

Table 10 Overall confusion matrix achieved in the real-world experiment [23]

Class:	Classified as:					
	Broken glass	Gunshot	Scream	Other	Precision	Recall
Broken glass	105	1	1	3	0.739	0.955
Gunshot	33	326	0	3	0.906	0.901
Scream	4	3	179	7	0.994	0.803
Correct classifications / all events (accuracy)	610/695	(87.77 %)				

For the localization technique considered, the accuracy was strongly connected to the SNR value. Its accuracy was high for SNR greater than 15 dB for impulsive sounds events and for SNR greater than 5 dB for scream cases. Moreover, the type of disturbing noise also had a principal influence on the results obtained. Traffic noise had the lowest impact on localization precision as opposed to indoor noise. The application of other digital signal processing techniques, such as band pass or recursive filtration, can significantly increase the accuracy of the sound source localization. Another essential improvement for localization, especially for impulsive sounds, could be made by changing the frame length. The frame length used, of about 85 ms, could be too wide for impulsive sound events, whereas such a frame length was appropriate for scream events.

In a related work the aspect of decision making time was investigated [24]. In a practical automatic surveillance system the latency is very important. It was shown that owing to parallel processing, the time needed to make the decision can be reduced to approximately 100 ms. Such a value is comparable with the so-called low-latency audio applications. One of the key findings of this related article is that the algorithms introduced in that work are capable of very fast online operation.

To summarize, the research has proved that the engineered methods for recognizing and localizing acoustic events are capable of operating in noisy conditions with moderate noise levels preserving an adequate accuracy. It is possible to implement the methods in an environmental audio surveillance system, working in both indoor and outdoor conditions. The proposed novel detection algorithms are able to robustly detect events even with SNRs below 0. As expected, the classification of acoustic events is more prone to errors in the presence of noise. However, some events are still accurately recognized at low SNRs.

Table 11 Comparison of recall and precision rates achieved in the anechoic chamber and in the real-world experiment

Event	Precision	Recall
Broken glass (real)	0.762	0.889
Broken glass (anechoic)	0.568	0.885
Gunshot (real)	0.796	0.915
Gunshot (anechoic)	0.8	0.815
Scream (real)	1	0.622
Scream (anechoic)	0.774	0.571

Acknowledgments Research is subsidized by the European Commission within FP7 project "INDECT" (Grant Agreement No. 218086). The presented work has been also co-financed by the European Regional Development Fund under the Innovative Economy Operational Programme, INSIGMA project no. POIG.01.01.02-00-062/09.

References

1. Basten T, de Bree H.-E., Druyvesteyn E et al. (2009) Multiple incoherent sound source localization using a single vector sensor ICSV16, Krakow, Poland
2. Basten T, de Bree H.-E, Tijs E et al. (2007) "Localization and tracking of aircraft with ground based 3D sound probes". 33rd Europ Rotorcraft Forum, Kazan
3. Chang CC, Lin CJ (2011) "LIBSVM: A library for support vector machines,". ACM Trans Intell Syst Technol (TIST), 2, 3, article 27
4. Cowling M, Sitte R (2003) Comparison of techniques for environmental sound recognition". Pattern Recogn Lett 24:2895–2907
5. Dat T, Li H (2010) Sound event recognition with probabilistic distance SVMs". IEEE Trans Audio Speech Language Process 19(6):1556–1568
6. de Bree H-E (2003) The Microflown: an acoustic particle velocity sensor. Acoust Aust 31(3):91–94
7. de Bree DH, Druyvesteyn WF (2005) "A particle velocity sensor to measure the sound from a structure in the presence of background noise,". Proc Int Conf FORUM ACUSTICUM
8. Dennis J, Tran H, Chng E (2013) Overlapping sound event recognition using local spectrogram features and the generalised hough transform. Pattern Recogn Lett 34(9):1085–1093
9. Donzier A, Cadavid S (2005) Small arm fire acoustic detection and localization systems: gunfire detection system, Proc. SPIE 5778, sensors, and command, control, communications, and intelligence (C3I) technologies for homeland security and homeland defense IV, 245 doi:10.1117/12.607128;
10. George J, Kaplan LM (2011) Shooter localization using soldier-worn gunfire detection systems, 14th International Conference on Information Fusion Chicago, Illinois, USA
11. Hearst MA (1998) Support vector machines. IEEE Intell Syst Their Applic 13(4):18–28
12. Jacobsen F, de Bree HE (2005) A comparison of two different sound intensity measurement principles". J Acoust Soc Am 118(3):1510–1517
13. Kiktova-Vozarikova E, Juhar J, Cizmar A et al. (2013) Feature selection for acoustic events detection. Multimed Tools Applic. published online
14. Kim H-G, Moreau N, Sikora T (2004) Audio classification based on MPEG-7 spectral basis representations. IEEE Trans Circ Syst Video Technol 14(5):716–725
15. Kotus J (2010) Application of passive acoustic radar to automatic localization, tracking and classification of sound sources". Inform Technol 18:111–116
16. Kotus J (2013) "Multiple sound sources localization in free field using acoustic vector sensor". Multimed Tools Applic. published online doi: 10.1007/s11042-013-1549-y
17. Kotus J, Łopatka K, Czyżewski A. et al. Processing of acoustical data in a multimodal bank operating room surveillance system. Multimed Tools Appl. doi: 10.1007/s11042-014-2264-z
18. Kotus J, Łopatka K, Kopaczewski K et al. (2010) "Automatic audio-visual threat detection". IEEE Int Conf Multimed Commun, Services Security (MCSS 2010) 140–144, Krakow
19. Kotus J, Lopatka K, Czyzewski A et al. (2011) "Detection and localization of selected acoustic events in 3D acoustic field for smart surveillance applications". 4th Int Conf Multimed Commun, Services Security (MCSS 2011) 55–63, Krakow
20. Kotus J, Lopatka K, Czyzewski A (2014) Detection and localization of selected acoustic events in acoustic field for smart surveillance applications. Multimed Tools Appl 68:5–21

21. Krijnders JD, Niessen ME, Andringa TC (2010) Sound event recognition through expectancy based evaluation of signal-driven hypotheses. Pattern Recogn Lett 31:1552–1559
22. Lojka M, Pleva M, Juhar J et al. (2013) Modification of widely used feature vectors for real-time acoustic events detection. Proc 55th Int Symp. Elmar 199-202
23. Łopatka K, Czyżewski A. "Recognition of hazardous acoustic events employing parallel processing on a supercomputing cluster". 138th Audio Eng Soc Convention 7-10.5.2015, Warsaw
24. Łopatka K, Czyżewski A (2014) Acceleration of decision making in sound event recognition employing supercomputing cluster. Inf Sci 285:223–236
25. Łopatka K, Żwan P, Czyżewski A (2010) Dangerous sound event recognition using support vector machine classifiers. Adv Intell Soft Comput 80:49–57
26. Lu L, Zhang H, Jiang H (2002) Content analysis for audio classification and segmentation". IEEE Trans Speech Audio Process 10(7):504–516
27. Machine Learning Group at University of Waikato (2012) "Waikato environment for knowledge analysis". http://www.cs.waikato.ac.nz/ml/weka
28. Mesaros A, Heittola T, Eronen A et al. (2010) "Acoustic event detection in real life recordings," 18th Europ Sig Process Conf 1267–1271
29. Millet J, Baligand B (2006) Latest achievements in gunfire detection systems. In battlefield acoustic sensing for ISR applications (26-1–26-14). Meeting Proc RTO-MP-SET-107, Paper 26. Neuilly-sur-Seine, France: RTO
30. Ntalampiras S, Potamitis I, Fakotakis N (2011) Probabilistic novelty detection for acoustic surveillance under real-world conditions. IEEE Trans Multimed 13(4):713–719
31. Ntalampiras S, Potamtis I, Fakotakis N (2009) "An adaptive framework for acoustic monitoring of potential hazards". EURASIP J Audio Speech Music Process 594103:1–15
32. Peeters G (2004) "A large set of audio features for sound description (similarity and classification) in the CUIDADO project", published online http://www.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf
33. Platt JC (1998) Sequential minimal optimization: a fast algorithm for training support vector machines. Adv Kernel Methods, Support Vector Learn 208(14):1–21
34. Raangs R, Druyvesteyn WF (2002) Sound source localization using sound intensity measured by a three dimensional PU probe, AES Munich
35. Rabaoui A, Davy M, Rossignol S, Ellouze N (2008) Using one-class SVMs and wavelets for audio surveillance. IEEE Trans Inform Forensics Sec 3(4):763–775
36. Rabaoui A, Kadri H, Lachiri Z et al. (2008) "Using robust features with multi-class SVMs to classify noisy sounds". 3rd Int Symp Commun, Control Sig Process 594–599 Malta
37. Raytheon BBN Technologies, "Boomerang", http://www.bbn.com/boomerang
38. Safety Dynamics Systems, "SENTRI", http://www.safetydynamics.net
39. SST Inc., "ShotSpotter", http://www.shotspotter.com
40. Temko A, Nadeu C (2009) Acoustic event detection in meeting room environments". Pattern Recogn Lett 30:1281–1288
41. Tijs E, de Bree H.-E, Steltenpool S et al. (2010) "Scan & Paint: a novel sound visualization technique". Inter-Noise 2010, Lisbon
42. Valenzise G, Gerosa L, Tagliasacchi M et al. (2007) "Scream and gunshot detection and localization for audio-surveillance systems". Proc IEEE Conf Adv Video Sig Based Surveill, London 21–26
43. Wind JW (2009) Acoustic source localization, exploring theory and practice. PhD The-sis, University of Twente, Enschede, The Netherlands
44. Wind JW, Tijs E, de Bree H-E (2009) Source localization using acoustic vector sensors, a MUSIC approach. NOVEM, Oxford
45. Yoo I, Yook D (2009) Robust voice activity detection using the spectral peaks of vowel sounds". J Electron Telecommun Res Institute 31:451–453
46. Zhuang X, Zhou X, Hasegawa-Johnson M, Huang T (2010) Real-world acoustic event detection". Pattern Recogn Lett 31:1543–1551
47. Żwan P, Czyżewski A (2010) Verification of the parameterization methods in the context of automatic recognition of sounds related to danger". J Digit Forensic Pract 3(1):33–45

Kuba Łopatka graduated from Gdansk University of Technology in 2009, majoring in sound and vision engineering. He completed his doctoral studies in 2013 at the Multimedia Systems Department and, at the moment of the submission of this article, works on completing his PhD dissertation on detection and classification of hazardous acoustic events. His scientific interest lies in audio, signal processing, speech acoustics and pattern recognition. He is an author or co-author of over 30 published papers, including 4 articles in journals from the ISI master journal list. He has taken part in various research projects, concerning intelligent surveillance, multimodal interfaces and sound processing.

Dr. Jozef Kotus graduated from the Faculty of Electronics Telecommunications and Informatics, Gdansk University of Technology in 2001. In 2008 he completed his Ph.D. under the supervision of prof. Bożena Kostek. His Ph.D. work concerned issues connected with application of information technology to the noise monitoring and prevention of the noise-induced hearing loss. He is a member of the international organization of the Audio Engineering Society (AES) and European Acoustics Association (EAA). Until now he is an author and co-author more than 50 scientific publications, including 11 articles from the ISI Master Journal List and 32 articles in reviewed papers. Also 3 chapters of books published by Springer were issued. He has extensive experience in sound and image processing algorithms.

Prof. Andrzej Czyzewski - Head of the Multimedia Systems Department is author of more than 400 scientific papers in international journals and conference proceedings. He has led more than 30 R&D projects funded by the Polish Government and participated in 5 European projects. He is also author of 8 Polish patents and 4 international patents. He has extensive experience in soft computing algorithms and sound & image processing for applications among others in surveillance.