Processing of acoustical data in a multimodal bank operating room surveillance system

J. Kotus · K. Łopatka · A. Czyżewski · G. Bogdanis

Abstract An automatic surveillance system capable of detecting, classifying and localizing acoustic events in a bank operating room is presented. Algorithms for detection and classification of abnormal acoustic events, such as screams or gunshots are introduced. Two types of detectors are employed to detect impulsive sounds and vocal activity. A Support Vector Machine (SVM) classifier is used to discern between the different classes of acoustic events. The methods for calculating the direction of coming sound employing an acoustic vector sensor are presented. The localization is achieved by calculating the DOA (Direction of Arrival) histogram. The evaluation of the system based on experiments conducted in a real bank operating room is presented. The results of sound event detection, classification and localization are provided and discussed. The practical usability of the engineered methods is underlined by presenting the results of analyzing a staged robbery situation.

Keywords Sound event detection · Sound source localization · Audio surveillance · Acoustic vector sensor

1 Introduction

Owing to the recent development of automatic visual and acoustic event detection methods, practical applications of audio-visual surveillance solutions are now possible. In our work we evaluate an application of such a technology in a bank operating hall. Over the years CCTV has become the elementary technical surveillance system for banks. The more and more

J. Kotus (✉) · K. Łopatka · A. Czyżewski
Faculty of Electronics, Telecommunications and Informatics, Multimedia Systems Department, Gdańsk University of Technology, Gdańsk, Poland
e-mail: joseph@sound.eti.pg.gda.pl

K. Łopatka
e-mail: klopatka@multimed.org

A. Czyżewski
e-mail: andcz@multimed.org

G. Bogdanis
Informatic systems designing and applications agency Microsystem, Sopot, Poland

sophisticated algorithms for video image analysis can complement or even replace the existing Intruder and Hold Up Alarm Systems (I& HAS).

During the year 2012 there were above 1,000 crimes committed on the premises of banks across Poland. These were either offences directed at ATMs, bank raids with weapons, or objects imitating weapons, and thefts taking place on bank grounds. Only on two occasions there were attempts to break into the bank and it was twice that the bank convoy was attacked.

Undoubtedly, a large number of bank raids were registered via CCTV and such material was further used by the dedicated investigative services for recreating the course of events and initiating the search for the assumed perpetrators. The collected recordings enable capturing of up to 50 % (depending on the region of the country) of the offenders. In our work we aim at detecting acoustic events in the bank environment, calculating the location of the sound sources and discerning between the threatening and non-threatening events. The hazardous events are, e.g.. screams or gunshots. The employed localization technique provides the information about the location of the sound source, which is a practically useful feature. The knowledge of the location of the event (acoustic direction of arrival) can be used to improve the efficiency of security surveillance, e.g. by automatically pointing the PTZ camera towards the direction of the detected action.

According to the concept diagram presented in Fig. 1, the input signals originate from the multichannel acoustic vector sensor (AVS). Next, the signal processing is performed. Initially, the detection of acoustic events is carried out. Subsequently, the detected events are classified to determine whether the action poses any threat or not. Finally, the acoustic direction of arrival (DOA) is calculated using the multichannel signals from the AVS. The methods employed for detection and classification, as well as the localization algorithm, are described in the next sections.

The algorithms operate in realistic conditions, in the presence of disturbing noise and room reflections. In the following sections an attempt to assess the performance of the employed signal processing techniques in such difficult conditions is made. The results gathered from the analysis of both live audio data and recorded situations, including arranged threats, are presented. It is shown that the performance of the system is sufficient to detect alarming situations in the given conditions.

The paper is organized as follows. In Section 2 the algorithms employed for the task of automatic detection and classification of events are described. Next, the methodology of event localization is outlined. In Sec. 4 we present conducted experiments and their results. Finally, in Sec. 5 the conclusions are drawn.
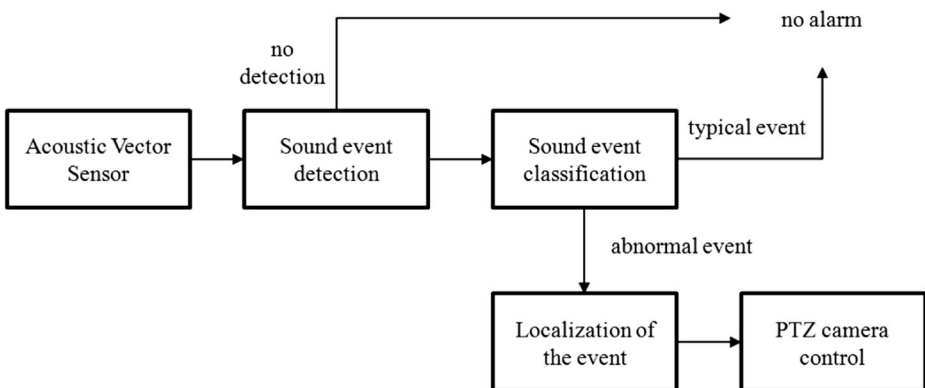

Fig. 1 Concept diagram of the audio-visual bank operating hall surveillance system

2 Acoustic event detection and classification

In the state-of-the art research on acoustic event recognition the most popular approach employs generative (e.g. Gaussian Mixture Models or Hidden Markov Models) or discriminative (e.g. Support Vector Machines, Artificial Neural Networks) pattern recognition algorithms to discern between the predefined types of sounds basing on the extraction of either: spectral, temporal or perceptual features [3–6]. Some works follow the so-called detection by classification approach, in which the classifier operates online and the decision concerning the foreground acoustic event presence is made basing on the results of classification [4]. The other approach, denoted detection and classification, divides the process into two operations – detection is responsible for separating the foreground event from the acoustic background; classification – for recognizing the type of detected sound [7]. In our work we follow the second approach. It allows us to employ the threshold-based detection, which is particularly useful in this application. As it is shown in the following sections, the threshold adaptation enables adapting to the changes in the acoustic environment. Thus, the algorithms are able to operate efficiently both during peak hours and at night. The bank operating hall is an indoor acoustic environment in which most sounds are generated by people and comprise:

– background cocktail-party noise,
– foreground voice activity,
– stamping,
– other sounds (e.g. chairs moving with a squeak, safes beeping, people's steps, objects being put down on desks etc.).

These sounds are regarded as typical elements of the acoustic environment. The abnormal sounds considered are gunshots and screams. We define the following classes of acoustic events in order to discern between them: speech, scream, gunshot, stamp, chair, beep and other. The spectrograms of example acoustic events encountered in the bank operating hall are depicted in Fig. 2. A 42 ms frame was used with 5 ms step. The frequency resolution of the spectrogram equals 23.4 Hz. Differences in the spectral and temporal characteristics of sounds are discernible quite easily. These differences are reflected by signal features and enable the distinction between the different event types. To detect and to recognize the events we use two types of detectors and one classification algorithm, which will be introduced in the following subsections.

2.1 Detection

Two types of detectors are employed. Impulse detector is designed to detect short impulsive sounds (e.g. stamps) [2]. Speech detector is intended to detect tonal sounds, and voice activity in particular [1]. The impulse detector algorithm is based on comparing the instantaneous equivalent sound level L with the threshold t. The sound level is calculated according to the formula as in Eq. (1):

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

where $x[n]$ represents the samples of the analyzed digital signal and $N = 512$ samples (at 48,000 samples per second) equals the length of the analysis frame. The parameter L_{norm}
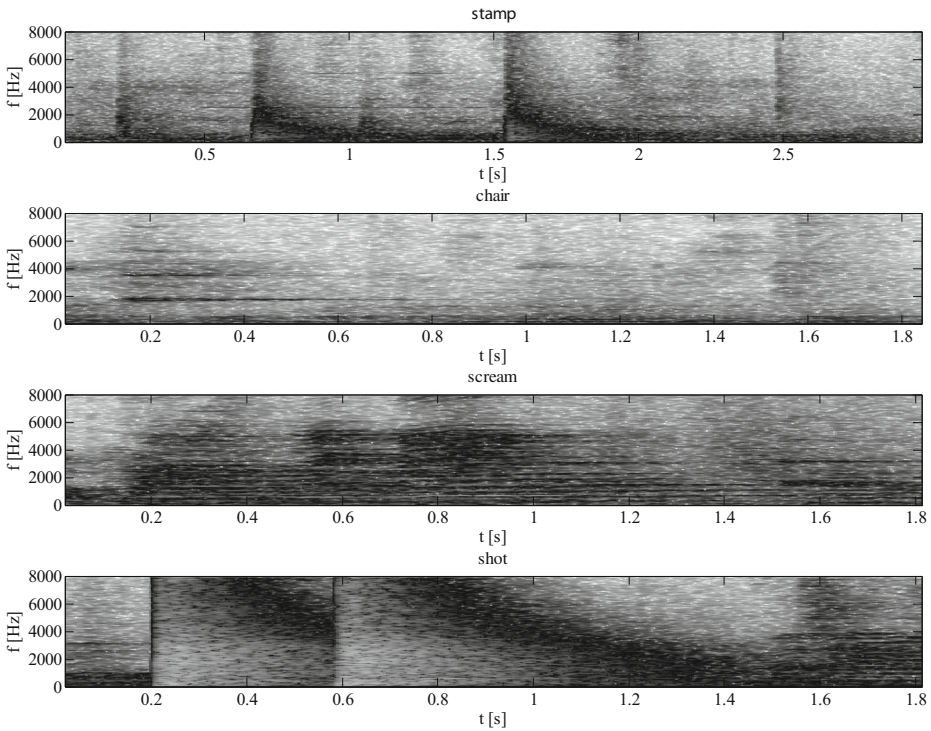
Fig. 2 Spectrograms of example sound events of selected type

(normalization level) assures that the result is expressed in dB SPL (relative to 20 μPa). The speech detector is based on the parameter peak-valley difference (PVD) defined in Eq. (2) [8]:

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

The PVD parameter is calculated from the signal power spectrum $X(k)$ calculated in 4,096 sample frames. The vector P contains locations of the spectral peaks. In order to calculate the parameter value, locations of the spectral peaks must be known. Herewith, we employed a simple grid search algorithm capable of finding the spectral peaks. The detection threshold t is obtained by adding a 10 dB margin to the average sound level in case of impulse detection or multiplying the median PVD value by 2 in case of speech detector. The threshold is smoothed using exponential averaging according to Eq. (3):

$$t = t_{old} \cdot (1-\alpha) + t_{new} \cdot \alpha \tag{3}$$

The exponential averaging enables an adaptation of the detection algorithm. The new threshold value t_{new} utilized in Eq. 3 is introduced in order to consider the changes in the acoustic background. The constant α is related to the detector's adaptation time and is obtained

from the formula in Eq. (4). In this case T_{adapt} equals an assumed value of 10 min, which yields α equal to $1.8 \cdot 10^{-5}$:

$$T_{adapt}[s] = \frac{N}{SR \cdot \alpha} \qquad (4)$$

where SR denotes the sampling rate – in this case 48,000 Sa/s and N denotes the length of the detector frame – 512 samples for the impulse detector and 4,096 samples for the speech detector.

2.2 Classification

A Support Vector Machine classifier (SVM) was used to classify the events. The seven classes of events (*speech, scream, gunshot, stamp, chair, beep, other*) are recognized. Therefore, a multiclass SVM with 1-vs-all approach was employed. The polynomial kernel function with degree 5 was used. It was found in preliminary experiments that the polynomial kernel outperforms the Radial Basis Function (RBF) and the sigmoid kernel as far as this classification task is concerned. For the training of the classification model, the γ parameter was set to 0.01 and the cost parameter C was set to 0.1. These parameters yielded the best performance on the training set.

A set of 55 features was used for the classification. These include *Audio Spectrum Envelope* (ASE), *Audio Spectrum Spread* (ASS), *Audio Spectrum Centroid* (ASC), *Zero Crossing Density* (ZCD), *Spectral Flatness Measure* (SFM), *Peak-Valley-Difference* (PVD), *periodicity*, f_0, *Spectral Energy*, transient parameters and *MFCCs*. According to the authors' experience, spectral shape and temporal features are essential for discerning between different types of acoustic events. The basic spectral shape parameters, such as ASE, ASC or ASS, highlight the important differences between the acoustic representations of the considered sound events. A gunshot, e.g. has a much wider spectrum than a scream event, which is reflected by the ASS parameter. To recognize scream from regular speech, the values of the ASC and f_0 features can be considered, since screams have their energy concentrated in higher regions of the spectrum. Furthermore, the spectral flatness feature (SFM), or the PVD parameter defined in Eq. (2) reflect the tonality of vocal sounds as opposed to other classes of events, which can be considered atonal. Further details concerning classification and calculation of the features can be found in related publications [9–11].

The employed classification algorithm does not support overlapping events. However, as it is shown in some examples presented later in this paper, events can still be correctly recognized if they are very close in time to each other, still only one event can be recognized in one time frame.

3 Acoustic events localization

Acoustic vector sensors were first applied to acoustic source localization in the air by Raangs et al. in 2002, who used the measured sound intensity vector to localize a single monopole source [12]. 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 (MUltiple SIgnal Classification) method to localize up to two sources using a single acoustic vector sensor [13]. In the same year Wind et al. applied the same method to localize up to four sources using two acoustic vector sensors [14, 15].

The authors' experiences with the sound source localization based on sound intensity methods performed in the time domain or in the frequency domain were presented in our previous papers [16–18]. The developed algorithm used for acoustic events localization operates in the frequency domain. Its block diagram is depicted in Fig. 3. The multichannel
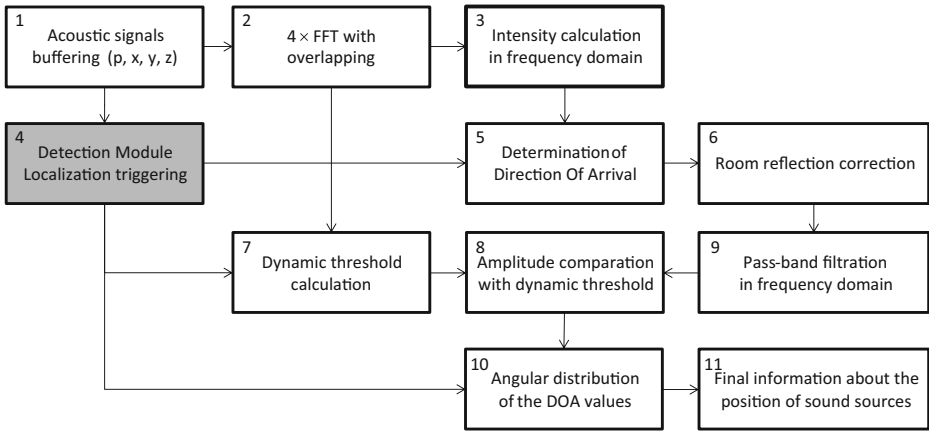
Fig. 3 The block diagram of the proposed algorithm

acoustic vector sensor produces the following signals: sound pressure p and three orthogonal particle velocity components u_x, u_y, u_z. The essential functionality of the localization algorithm is its connection with the acoustic event detection module (described in details in section 2.1). The detection module (block 4 indicated by gray shading) operates on acoustic pressure signal only. The detection procedure is performed in parallel to the localization process. Signal set by the detection module affects the operation of the units 5, 7 and 10.

In block 2 the acoustical signals are buffered and prepared for FFT (Fast Fourier Transform) calculation. The Hanning window is applied. Subsequently, the 4,096 point FFT for each signal is performed, with the sampling rate equal to 48 kS/s (frequency resolution: 11.7 Hz). Such parameters provide a sufficient spectral resolution for sound source localization. The overlap factor between frames is equal to 50 %. The FFT calculation is performed for each acoustic component (p, u_x, u_y, u_z), separately. This operation yields transformed signals: $X_p[i]$, $X_{ux}[i]$, $X_{uy}[i]$, $X_{uz}[i]$, where i (ranging from 0 to 4,095) denotes the index of the spectral bin. The matrix X now contains information about the direction of arrival of every spectral component of the signal.

$$\mathbf{X} = [\,\mathbf{X}_p \quad \mathbf{X}_{ux} \quad \mathbf{X}_{uy} \quad \mathbf{X}_{uz}\,] = [\,\Im\{p[n]\} \quad \Im\{u_x[n]\} \quad \Im\{u_y[n]\} \quad \Im\{u_z[n]\}\,] \qquad (5)$$

To extract the information about DOA from the signals in the frequency domain, the sound intensity in the frequency domain is computed (block 3 of Fig. 3). The sound intensity vector is defined and calculated according to Eq. (6),

$$\overrightarrow{I}[i] = \begin{bmatrix} I_x[i] \\ I_y[i] \\ I_z[i] \end{bmatrix} = \begin{bmatrix} X_p[i]\cdot X_{ux}[\bar{i}] \\ X_p[i]\cdot X_{uy}[\bar{i}] \\ X_p[i]\cdot X_{uz}[\bar{i}] \end{bmatrix} \qquad (6)$$

where:

$I_x[i]$ sound intensity component for x direction for i-th spectral bin
$\underline{X_p[i]}$ coefficients of complex spectrum for i-th spectral bin for acoustic pressure signal
$\overline{X_{ux}[i]}$ conjugated spectrum coefficients for particle velocity in x direction.

The calculation of sound intensity components in the frequency domain is the most important part of the localization algorithm. These data are used in block 5. Before the calculation of the direction of arrival for particular frequency partials, the signal from the detection module is included in the calculation.

The positive state of the detection module initiates the following actions:

– the dynamic threshold calculation in the frequency domain is suspended (block 7),
– time distribution of direction of arrival (DOA) is prepared to be calculated (block 10),
– the calculation process of DOA for detected sound event begins (block 5).

In block 5 the angular values for azimuth and elevation for i-th spectral components are computed according to Eq. (7):

$$
\begin{cases}
r[i] = \sqrt{I_x[i]^2 + I_y[i]^2 + I_z[i]^2} \\
\varphi[i] = \arctan\left(\dfrac{I_x[i]}{I_y[i]}\right) \\
\theta[i] = \arcsin\left(\dfrac{I_z[i]}{r[i]}\right)
\end{cases}
\tag{7}
$$

where $\phi[i]$ is the azimuth angle, $\theta[i]$ and is the elevation angle corresponding to the i-th spectral bin.

In the next step (block 6) the obtained angular values are corrected according to specular reflection [19]. This kind of correction is applied for azimuth angle values from the range: 0 to 180° (reflections produced by the wall behind the sensor) and for elevation angle values from the range: 0 to –90 (reflections produced by the ceiling above the sensor) - see the Fig. 6 for details.

The most of the acoustic energy for the speech signal is contained in the frequency range from about 200 to 4,000 Hz. In case of the considered speech signal, for a better separation from other disturbing components, an additional filtration in the frequency domain is applied, which is handled in block 9. At the output of this block the DOA values are limited to the frequency range from 200 to 4,000 Hz. After that the next step of the localization algorithm can be executed, within the block 8. In such a case the data from blocks 7 and 9 are used. For a better understanding this step the explanation of block 7 is provided below. The computation of the dynamic threshold is performed for all frequency components in the FFT spectrum independently, according to Eq. (8):

$$
\left|X_{p\_th}[i]\right| = \left|X_{p\_old}[i]\right| \cdot (1-\alpha) + \left|X_{p\_new}[i]\right| \cdot \alpha
\tag{8}
$$

where: $|X_{p\_th}[i]|$ - is the magnitude of the acoustic pressure which will be used to calculate the dynamic threshold, i – component number of the FFT spectrum, indices *old* and *new* mean the earlier and the present value of $|X_p[i]|$, respectively and the constant α is related to the spectrum adaptation time and is in this case equal to 0.05.

The time averaging process is carried out when the detector module returns 0 (no event detected). As it was mentioned before, the averaging process is suspended when the event detection is triggered. In that case the final average values are used as the dynamic threshold.

In the block 8 the values from the block 9 are compared with the dynamic threshold obtained from the block 7 according to the condition expressed by formula (9):

$$L_{p\_event}[i] \geq L_{p\_th}[i] + 10 \tag{9}$$

where: L_{p_event} – sound pressure level for considered acoustic event, L_{p_th} – sound pressure level of dynamic threshold, i-th point number of the FFT spectrum.

If the condition expressed by Eq. (9) is true for a given FFT component, the DOA values computed in the block 9 (corrected and filtered azimuth and elevation values for particular FFT components) are used to compute the angular distribution of azimuth and elevation. This operation is performed in the block 10. It is important to emphasize that the azimuth and elevation distributions are prepared to be calculated every time a new acoustic event is detected (because the algorithm operates online, the data computed for the last event are erased and the new empty tables are prepared). Both angular distributions (for azimuth and elevation) are computed using the sound intensity value for given angular values according to the formula.

$$\begin{cases} F_A[\varphi] = \sum I_\varphi \\ F_E[\theta] = \sum I_\theta \end{cases} \tag{10}$$

where $F_A[\phi]$ is the azimuth distribution, $F_E[\theta]$ is the elevation distribution, I_ϕ – the intensity value for given azimuth angle, I_θ – the intensity value for given elevation angle.

Finally, the values stored in the tables $F_A[\phi]$ and $F_E[\theta]$ are smoothed by means of weighted moving average and Hanning window coefficients (window length was equal to 5), according to (11):

$$\begin{cases} F_{SA}[\varphi] = \sum_{n=-2}^{n=2} F_A[\varphi + i] \cdot h_n & F_{SE}[\theta] = \sum_{n=-2}^{n=2} F_E[\theta + i] \cdot h_n \end{cases} \tag{11}$$

where: $F_{SA}[\phi]$ is the smoothed azimuth distribution, $F_{SE}[\theta]$ is the smoothed elevation distribution, i – index number, h_n – Hanning window coefficients (window length was equal to 5 frames).

The maximum values of the tables $F_{SA}[\phi]$ and $F_{SE}[\theta]$ indicate the current position of the considered acoustic event.

4 Results

4.1 Detection and classification results

First, an experiment was conducted in order to assess the detection and classification of typical acoustic events. A 13-hour-long signal registered in the bank operating hall was analyzed (recorded from 7:00 a.m. to 8:00 p.m.). The ground truth data concerning the position of sound sources were unknown. The detailed results of the sound source localization were presented in Section 4.2. The results of event detection are presented in Table 1. In total 1,026 acoustic

Table 1 Number of detected events

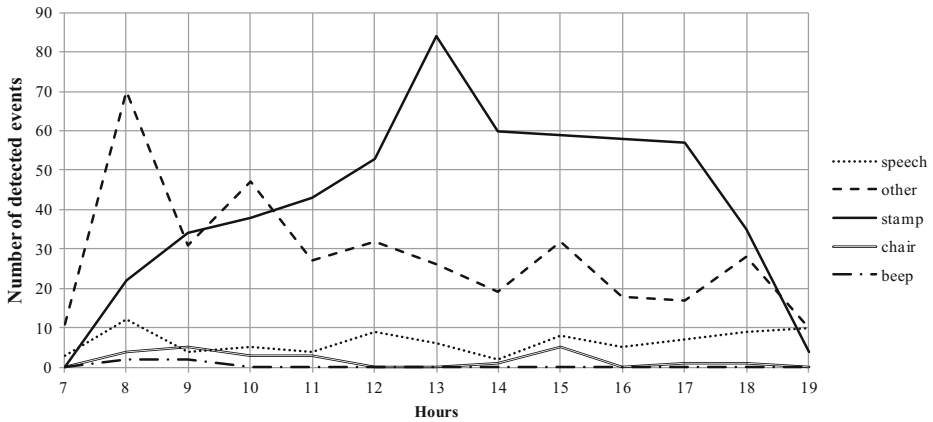| Type of event:	Speech	Rother	Stamp	Chair	Beep
Number of detecteed events:	84	368	547	23	4
				Total:	1,026

Fig. 4 Distribution of detected events per operating hour

events were detected. It is visible that stamps (547 detections) are the prevailing sounds. Other events are also frequently present (368 detections). Only 84 speech events were recognized, which can be explained by the fact that speech is often present in the background as cocktail-party noise and the detectors adapt to that background by adjusting the detection threshold according to Eqs. 3 and 4. The sound of moving chairs and money safe beeps were seldom detected (23 and 4 occurrences, respectively). No threatening sound events were detected during that period, which means that the system did not produce a false alarm.

In Fig. 4 the distribution of events per hour of operation is presented. It is visible that during the early hours – 7 to 9 a.m. – other sounds are more frequent than stamping. Stamping becomes more frequent during the operating hours, which are 9 a.m.to 6 p.m.. Information about the movement in the bank can also be derived from the event distribution. The peak hour appears to be at 1 p.m.. Speech events are evenly distributed, since they are caused by both: the clients and the staff of the bank, who are present during the whole period.
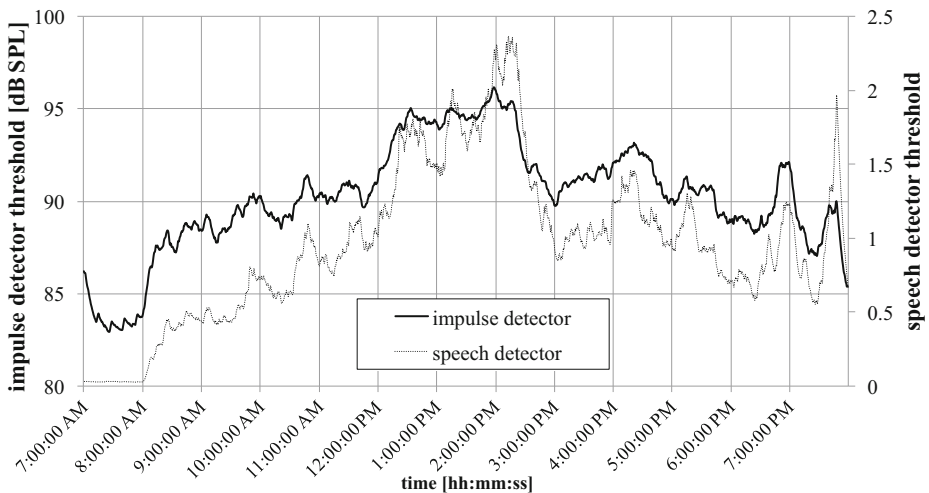

Fig. 5 Changes in detection threshold during operating hours

Table 2 Results of recognition of threatening events in an arranged situation

clasified as →	Speech	Scream	Gunshot	Other	Precision	Recall
Speech	24	7	0	10	1	0.58
Scream	0	21	0	4	0.75	0.84
Gunshot	0	0	4	0	1	1
Other	0	0	0	4	0.22	1
Overall correctly classified:					[53/74]	71.6 %

Another important aspect is the level of the acoustic background. As it was stated in the previous sections, the threshold of acoustic event detection is adaptive, i.e. it follows the changes of sound level in the environment. The changes of the detection threshold of both detection algorithms employed are presented in Fig. 5. Again, it is visible that the peak hour is around 1 p.m. to 2 p.m.. The SPL level of the background and the median PVD reach a maximum. During the early and the late hours the threshold is lowered, which enables the detection of some more subtle sound events.

To assess the efficiency of threatening events detection, a hazardous situation was arranged. 17 people took part in an experiment organized in a real bank operating hall. The arranged situations included shouting, screams, typical conversation and robbery simulation with the use of a gun. A noise gun was used to generate gunshot sounds. A 31-minute-long signal was registered and then analyzed. A total of 74 events were detected. Sounds belonging to the following classes were emitted: *speech, scream, gunshot* and *other*.

The analysis of the classification results is presented in Table 2. It can be observed that the threatening events are recognized with an acceptable accuracy. However, speech sounds are often confused with other events. All gunshots were correctly classified, although only four shots were emitted. It is worth noting that the events in this experiment were recognized in fully realistic conditions, namely in the presence of noise (mostly cocktail-party noise).

4.2 Acoustic events localization results

In the previous section the detection and classification results were presented. In this section the detection and localization functionality of the developed solution are shown. All experiments were conducted in the bank operating room. The ground truth data of sound source position were available in this case. First of all, the detection and localization tests for a single
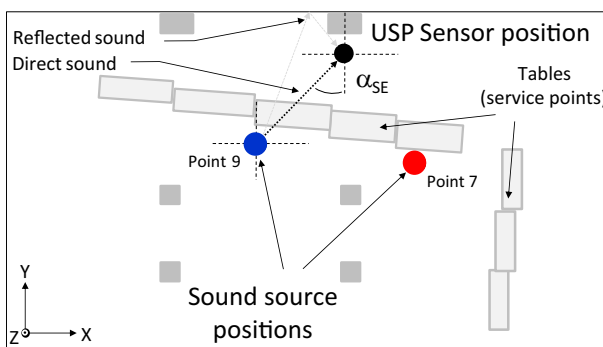
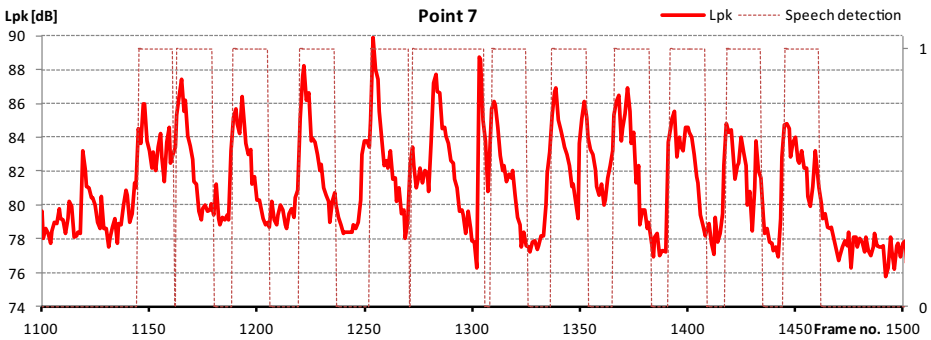**Fig. 6** Speaker positions during the sound source detection and localization experiments

Fig. 7 Speaker detection results obtained in point 7. This signal triggered the sound source localization algorithm

speaker were performed. The speaker was located successively in various customer service points (points 7 – red dot and 9 – blue dot in Fig. 6) and the spoken sequence consisted of counting from 1 to 10. In Fig. 6 the layout of the bank operating room is shown. The position of the USP sensor, direct sound and reflected sound paths are also illustrated. The group of gray rectangles indicates the position of customer service points (tables and chairs). The dimensions of the room were as follows: $x=20$ m, $y=15$ m, $z=3.5$ m. The experiments were conducted with the participation of 20 persons.

The prepared algorithm detected the spoken sentence, for which both azimuth and elevation values were determined.

The utterances were repeated three times for each particular point during this experiment. As it was mentioned in Section 3, the localization process is based only on the parts of the signal in which sound events are detected by the detection module (see the detection algorithm – section 2.1 and localization algorithm in section 3). In Fig. 7 the speaker detection for point 7 results are depicted. Results obtain in point 9 were shown in Fig. 8.

Two kinds of computation were done. First, the room reflection correction was disabled. In the second analysis the room reflections were considered. The sample results represented by the angular distributions $F_{SA}[\phi]$ and $F_{SE}[\theta]$ (for azimuth and elevation) are presented in Fig. 9 (without reflection correction) and in Fig. 10 (with reflection correction). The ground truth data are also depicted. The red dot indicates the point 7 and the blue diamond - point 9.
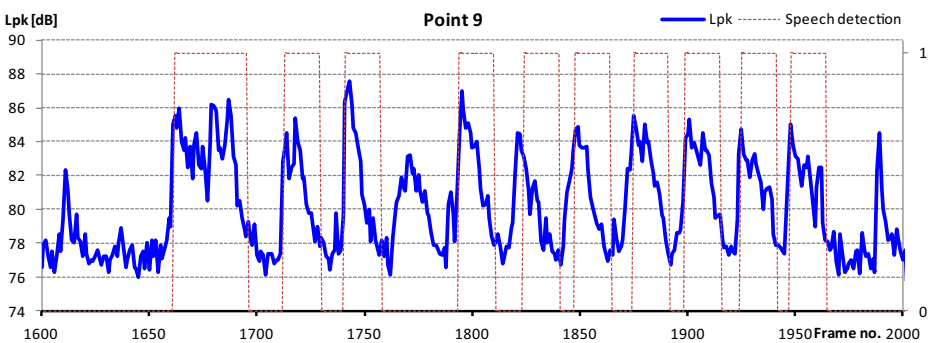

Fig. 8 Speaker detection results obtained in point 9. This signal triggered the sound source localization algorithm
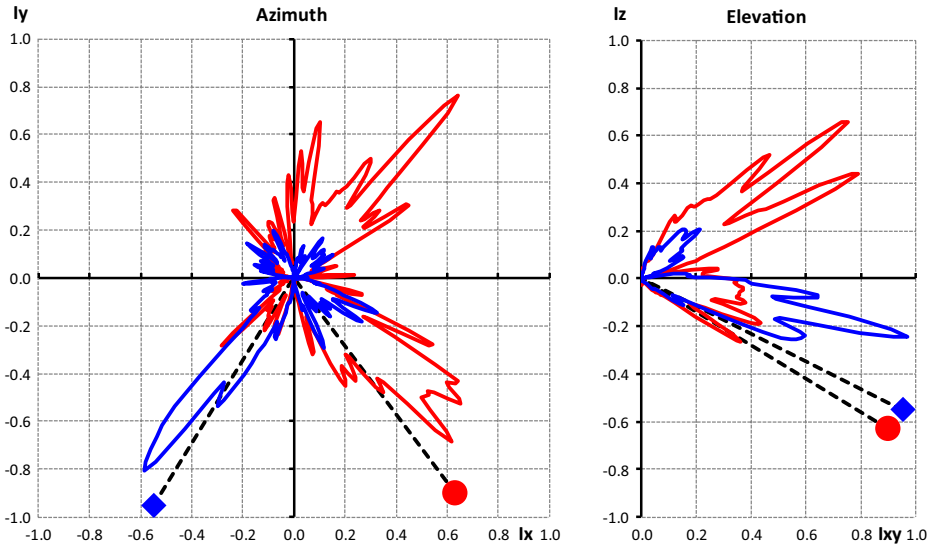
Fig. 9 Sample angular distributions for both: azimuth and elevation obtained for two speaker positions; without reflection correction

The localization results for the same signals, however analyzed with reflection correction, are shown in Fig. 10. Imaginary sound sources are reduced, so that the proper position of the considered sound sources is indicated. It proves that the reflection correction is an essential part of the localization algorithm.

All localization results for every detected sentence are presented in Fig. 11. The average values of azimuth and elevation for considered speaker positions were calculated. They are presented in Fig. 11. The red empty circle corresponds to the position 7, the blue empty diamond indicates the position 9. It is important to emphasize that the obtained average angle values are very close to the reference points.

The average values of azimuth and elevation calculated for all detected sound events spoken in both customer service points are shown in Table 3. Reference angular values are also collected. All angular values are expressed in degrees. Table 3 also contains the estimated
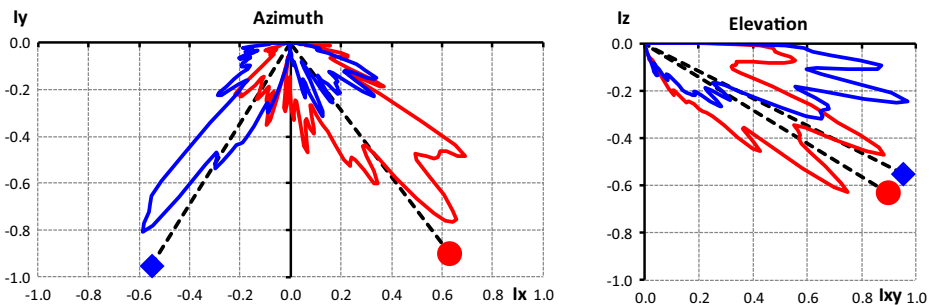

Fig. 10 Sample angular distributions both for azimuth and elevation obtained for two speaker positions; with reflection correction
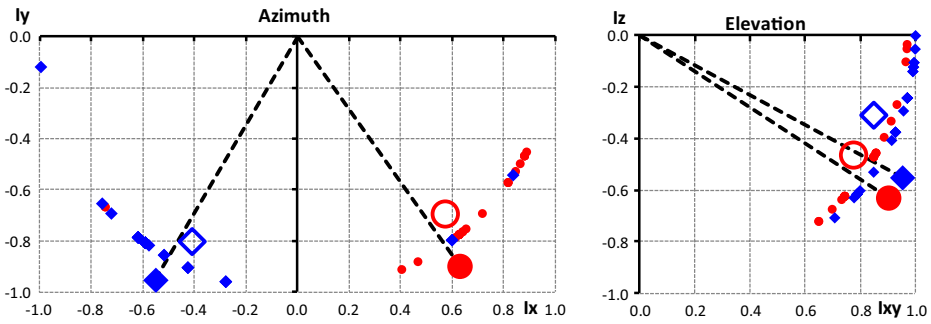
Fig. 11 Localization results expressed by the azimuth and elevation values for all detected sound events spoken in considered points

localization error (defined as a difference between the ground truth data and the calculated angles) for both settings: with (r.c. on) and without reflection correction (r.c. off).

4.3 Robbery event reconstruction

To underline the practical usability of the developed technology we present the results of the analysis of an arranged robbery event. The situation was performed in cooperation with security specialists. Scream and gunshot events were generated. A lot of commotion was caused, which resulted in an increased sound level throughout the event. The results presented in this section should be treated as a practical demonstration of the technology attempt, aiming to informal evaluation, only. The amount of data gathered in this experiment was too low to provide a reliable analysis of recognition and localization results. Instead, we focus on the spatial distribution and the temporal consequences of the detected events. We focused on events which are likely to happen in case of an actual bank robbery i.e. scream and gunshot.

In the first figure (Fig. 12) the sound pressure level (in 125 ms time base) is shown together with event detection results. Two spikes are visible. These come from gunshots. The other sound events are impossible to distinguish from the background based on signal level only. Hence, the engineered speech detector is employed and it yields correct detections. The spatial layout of the detected events is shown in Fig. 13. In Fig. 13 localization results concerned only a part of the operation room where the robbery event took place (the dimensions of the operating room were described in section 4.2.). The azimuth and elevation angles are calculated by the localization algorithm. The knowledge of the height on which the AVS is mounted, enables precise determination of sound source location. In Fig. 13 the detected

Table 3 Average results of sound source localization for particular points (in degrees)

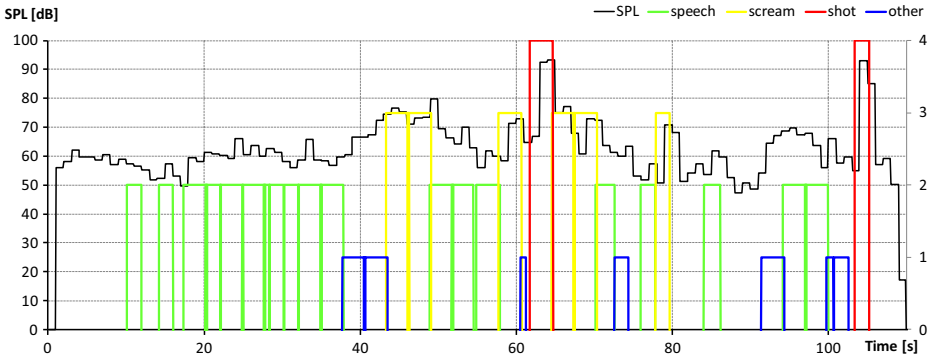| Source	Ground truth		Measurement results		Localization error	
Position	Azimuth	Elevation	Azimuth	Elevation	Azimuth	Elevation
P 7 r.c. on	300	35	309.5	30.8	9.5	4.2
P 9 r.c. on	245	30	243.0	19.9	2.0	10.1
P 7 r.c. off	300	35	203.9	−14.9	96.1	49.9
P 9 r.c. off	245	30	216.2	−0.8	28.8	30.8

Fig. 12 Sound pressure level (in 125 ms time base) and event detection results

events are drawn on the map of the room. The red dots indicate the position of the shot events. The green and yellow dots indicate the location of speech and scream events, respectively. Other sound events positions were marked using blue dots. In Fig. 14 the aggregation of detection, classification and localization results is presented. The changes of azimuth and elevation angle are visible when the sound source location changes (e.g. when another speaker takes over).

In Fig. 15 sample angular distributions for both: azimuth and elevation obtained for selected shot event (time: 63.14 [s]) are presented. Based on this data, the shooter position can be indicated. After the shots we observe scream events. The sample angular distributions for this event are presented in Fig. 16 (time: 65.38 [s]). These events were also produced by the attacker.
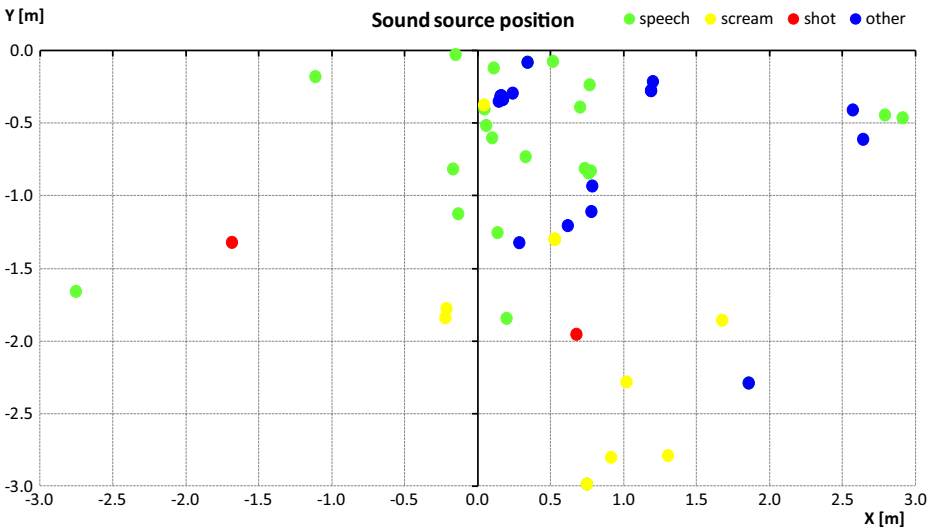

Fig. 13 Positions of sound sources in the bank operating room during robbery simulation. yellow dots indicate the scream events position, green dots are connected with speech events and blue dots indicate the other sound events
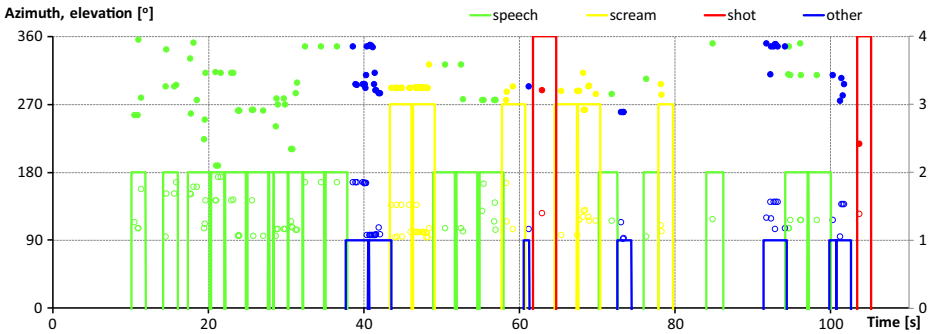
Fig. 14 Detection, classification and localization results for considered signal

5 Conclusions

The application of sound event detection, classification and localization methods to bank operating room surveillance was featured. The engineered algorithms were evaluated in practical conditions, in a real bank operating room, in the presence of typical background noise. The performance of the methods was assessed. It was shown that the automatic recognition of events enables a correct detection of both threatening and typical events. The localization algorithm also proved efficient enough to provide the information about the location of the sound sources in the considered space.

It was found on the basis of obtained results, that accurate acoustic event detection is required by the sound source localization algorithm to work properly. The designed and implemented algorithm of sound source localization was practically tested in real acoustic conditions. It was shown also, that the obtained average values of azimuth and elevation were close to the Ground Truth values. The error was less than 10° for the considered speaker positions.

It was shown that sound reflections inside the operating room can seriously disturb the localization process. The proposed room acoustics correction procedure ensures correct results, despite the presence of noticeable sound reflections.

The proposed methodology can significantly improve the functionality of the traditional surveillance monitoring systems. Application of acoustic vector sensor can be useful for localization of detected sound events. The described method can be applied to surveillance
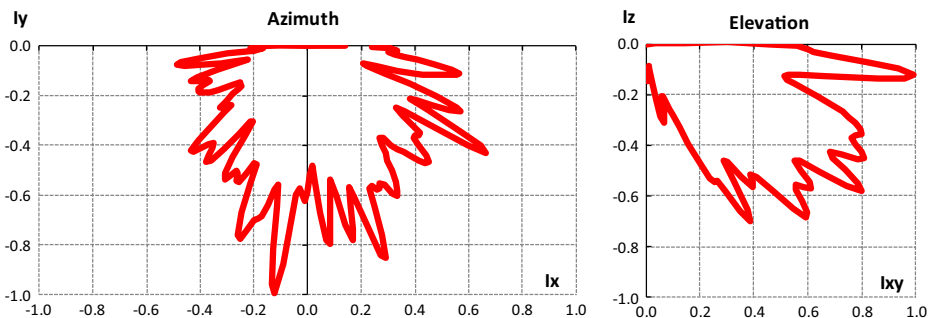

Fig. 15 Sample angular distributions for both: azimuth and elevation obtained for selected shot event (time: 63.14 [s])
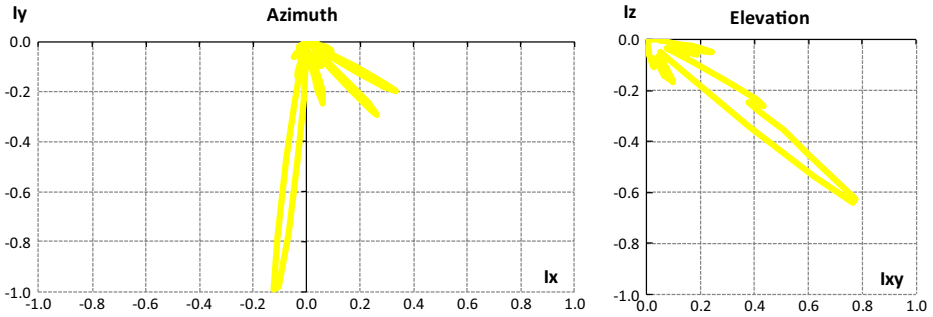
Fig. 16 Sample angular distributions for both: azimuth and elevation obtained for selected scream event (time: 65.38 [s])

systems for monitoring and visualising the acoustic field of a specified area. The direction of arrival information can be used to control the Pan-Tilt-Zoom (PTZ) camera to automatically point it towards the direction of the detected sound source. The presented event reconstruction technique can be used also to facilitate the work of the security personnel. The audio/video material from CCTV can be labelled according to the results of analysis. Thus, it is possible to automatically search for a relevant content in the stored data. Also, the proposed methods enable online visualization of the events (via the PTZ camera).

The integration of acoustic and visual event detection is also a promising technology. In a related work efforts were made to make the decision based on a joint analysis of audio and video streams [20]. In many cases of bank robberies the incident can be silent. Nevertheless, events such as the people putting their hands behind the head or kneeling on the floor can be detected by means of video analytics. Therefore, in the future work further multimodal analytics methods will be developed.

The acoustic modality offers many interesting functionalities in the automating detection and classification of hazardous situations. It is worth to emphasize that the proposed method can be a useful tool also during the offline forensic audio analysis.

References

1. Łopatka K, Kotus J, Czyżewski A (2011) Application of vector sensors to acoustic surveillance of a public interior space. Arch Acoust 36(4):851–860
2. Kotus J, Łopatka K, Czyżewski A (2012) Detection and localization of selected acoustic events in acoustic field for smart surveillance applications. Multimedia Tools Appl 07:2012
3. Cowling M, Sitte R (2003) Comparison of techniques for environmental sound recognition". Pattern Recogn Lett 24:2895–2907
4. Valenzise G, Gerosa L, Tagliasacchi M, Antonacci F, Sarti A (2007) Scream and gunshot detection and localization for audio-surveillance systems". IEEE Conf Adv Video Sig Based Surveill London 05–07(08): 21–26
5. Zhuang X, Zhou X, Hasegawa-Johnson M, Huang T (2010) Real-world acoustic event detection". Pattern Recogn Lett 31:1543–1551
6. A. Rabaoui, et al., "Using robust features with multi-class SVMs to classify noisy sounds", International Symposium on Communications, Control and Signal Processing, Malta, 12–14.03.2008
7. Temko A, Nadeu C (2009) Acoustic event detection in meeting room environments". Pattern Recogn Lett 30:1281–1288
8. Yoo I-C, Yook D (2009) Robust voice activity detection using the spectral peaks of vowel sounds. ETRI J 31(4):451–453
9. Łopatka K, Żwan P, Czyżewski A (2010) Dangerous sound event recognition using support vector machine classifiers. Adv Multimed Netw Inf Syst Technol 80:49–57
10. Zwan P, Czyzewski 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
11. Peeters G (2004) A large set of audio features for sound description (similarity and classification) in the cuidado project. CUIDADO IST Proj Rep 54(version 1.0):1–25
12. Raangs R. and Druyvesteyn W.F., 2002. Sound source localization using sound intensity measured by a three dimensional PU probe, AES Munich
13. Basten T, de Bree H-E, Druyvesteyn E (2009) Multiple incoherent sound source localization using a single vector sensor ICSV16. Krakow, Poland
14. Wind JW, Tijs E, de Bree H-E (2009) Source localization using acoustic vector sensors, a MUSIC approach. NOVEM, Oxford
15. Wind J.W., 2009. Acoustic Source Localization, Exploring Theory and Practice. PhD Thesis, University of Twente, Enschede, The Netherlands
16. Kotus J (2010) Application of passive acoustic radar to automatic localization, tracking and classification of sound sources. Inf Technol 18:111–116
17. Kotus J (2013) Multiple sound sources localization in free field using acoustic vector sensor. Multimedia Tools Appl(06)
18. Kotus J., Łopatka K., Czyżewski A., Bogdanis G., Audio-visual surveillance system for application in bank operating room, 6th International Conference on Multimedia, Communications, Services and Security, vol. 368, pp. 107 – 120, 6.6.2013 – 7.6.2013
19. Weyna S (2003) Identification of reflection and scattering effects in real acoustic flow field. Arch Acoust 28(3):191–203
20. Kotus J., Dalka P., Szczodrak M., Szwoch G., Szczuko P., Czyżewski A., Multimodal Surveillance Based Personal Protection System, Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) 2013, pp. 100 – 105, Poznań, Poland, 26.9.2013 – 28.9.2013

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 processing algorithms.

Mr. Kuba Lopatka is a Ph. D. student at the Multimedia Systems Department, Gdansk University of Technology. He graduated from the faculty of Electronics, Telecommunication and Informatics in 2009, majoring in sound and vision engineering and enrolled on the program of doctorate studies at the Multimedia Systems Department. The subject of his Ph. D. thesis is automatic recognition of sounds related to danger. So far he has published, as author or co-author, 26 publications including articles in international journals, 3 of which are noted on the ISI list, conference papers and chapters in books. His main scientific interests are analysis and processing of audio signals.

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.

Georgis Bogdanis graduated the faculty of Electronics, Telecommunication and Informatics in 1979 majoring in computer science and automatics. After graduating he worked for Gdańsk University of Technology. Since 1985, he is a President and Chief Executive of Microsystem Ltd. He is an expert in the field of security and alarm systems. Has contributed to various security-related projects, including works for gas and energy companies. With his company he has successfully introduced security systems in over 1,000 banks in Poland.