Gaze-tracking and acoustic vector sensors technologies for PTZ camera steering and acoustic event detection

Józef Kotus, Bartosz Kunka, Andrzej Czyżewski, Piotr Szczuko, Piotr Dalka, and Rafał Rybacki
Multimedia Systems Department
Gdansk University of Technology
Gdańsk, Poland
{joseph, kuneck, andcz, szczuko, dalken, rrybacki}@sound.eti.pg.gda.pl

***Abstract*—An innovative application of gaze-tracking and acoustic vector sensors (AVS) technologies for guidance of moving pan- tilt-zoom (PTZ) monitoring camera is presented. Gaze-tracking is used to steer and to zoom the camera to the gaze focus area. Additionally, it is combined with audio processing in two scenarios. First is called "audio slave": directional acoustic monitoring is adjusted automatically to the camera direction. Second is called "audio master": automatic detection of sound events directions is performed to take priority over user control and steer the camera towards sound source. An approach to gaze tracking is presented, utilizing new algorithmic methods for both image processing and PTZ camera steering. Then application of AVS for directional filtering of sound, and for detection of acoustic events direction is discussed. The implemented application is described, and user experience is reported. Finally, future work is discussed.**

Keywords-gaze-tracking; multimodal computer interfaces; sound processing, sound direction

I. INTRODUCTION

Current generation of monitoring systems provides access to digital video streams from multiple cameras. For the operator it is often overwhelming to control large number of cameras. It usually involves the use of numeric panel for camera selection and PTZ joystick for their steering. First the operator should recall the number of a particular camera, enter it, then switch to full screen view, engage manual operation mode and use PTZ joystick to guide the camera. We propose a hands-free, gaze-steered multi PTZ cameras system, where described operations are replaced by choosing camera by gaze focus on its image and to guide it automatically according to sight direction. Described setup can be also used by handicapped operators, or adopted to multimedia systems such as teleconferencing.

In the system a camera is equipped with innovative sound processing module, for acoustic directional monitoring, with sound reception direction matching the visual one. Moreover, the sound processing algorithm performs also audio event detection and direction estimation. In case of important event the user control is overridden and camera is automatically pointed at sound source.

II. GAZE-TRACKING FOR PTZ CAMERA STEERING

Gaze-tracking is a technique which allows tracking eye movements and estimating a localization of the sight fixation point, i.e. the point a user is looking at in the computer screen. It can be used for mouse cursor control by gaze and user's visual activity tests (in web page usability research or in concentration tests). Application described in this paper reveals new possibilities, particularly in visual monitoring employing moving cameras.

Up to date, the "eye-tracking" term is better known than "gaze-tracking". If the system in question is head-mounted and eye-in-head angles are measured, then one can refer to eye-tracking. Contrary, if the measurement camera is mounted on the computer monitor, then gaze angles are estimated. Next difference lies in the processing workflow: eye-tracking first estimates eye rotation angle, then take head position into account to calculate location of fixation point. Gaze-tracking estimates fixation point immediately in the first step, by determining how special IR markers reflect in the eye surface (Sec. A). There are many commercial gaze-tracking systems available on the market, usually very expensive [1]. Therefore the system engineered by the authors was used in described approach [2][3].

A. Gaze-tracking system setup

The developed gaze-tracking system is based on infrared (IR) illumination similarly to most of commercial systems. IR illumination is not visible to the user, therefore do not disturb interaction with the computer. The usage of IR light improves image registration significantly: contrast between the pupil and the iris in IR image is much higher compared to the standard gray-scale image taken utilizing only visible light illumination. Moreover, IR sources produce unique reflections on the eye, called glints (Sec. II.B), therefore increasing precision of gaze direction estimation.

The presented gaze-tracking system comprises of five major components (Fig. 1):

- modified webcam used to capture image of the user – camera sensitive to infrared, with customized lenses and infrared band-pass filter, mounted on a driver controlled by the software to follow user position, therefore it is not required for the user to sit still;
- IR LEDs ring placed on the camera;
- 4 groups of IR LEDs fixed on display corners;
- IR LEDs driver – the USB controlled device allowing separate activation of all mentioned IR modules;
- software for image processing and control of the IR LEDs driver, webcam driver and distant PTZ cameras steering [3].

The gaze-tracking system setup is presented in Fig. 2.

Differences in the iris colors among population and also lighting conditions influence the intensity of the IR reflection pattern (Fig. 3). The intensity adaptation for the IR is performed during calibration procedure performed before the system is ready for operation. During calibration the user is directed to look at nine predefined points sequentially presented on the display (in 3x3 grid). The algorithm compares the estimated fixation point with the known coordinates of the points on the screen. While processing the image of user's eyes, two operation modes are used interchangeably:

- *dark eye mode* – the IR LEDs on the camera are disabled and the intensity of IR emitted by the screen corners modules is increased (Fig. 3a),
- *bright eye mode* – the IR LEDs on the camera are activated to produce bright eye effect (Fig. 3b).

The dark eye mode is preferred when the user uses glasses. In this mode there is no light coming from the on-camera IR sources which may reflect from the glass and obscure the eye image. Generally, the dark eye mode provides less unwanted reflections on the eye. The bright eye mode is usually selected when the ambient light intensity is relatively low. According to the requirements of the standard IEC 60825-12:2004 the intensity of emitted IR light is low enough to be safe for the users eyes [5].

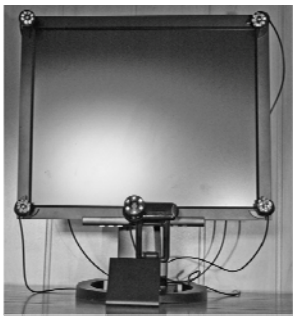

Figure 1. Hardware setup of the gaze-tracking system.

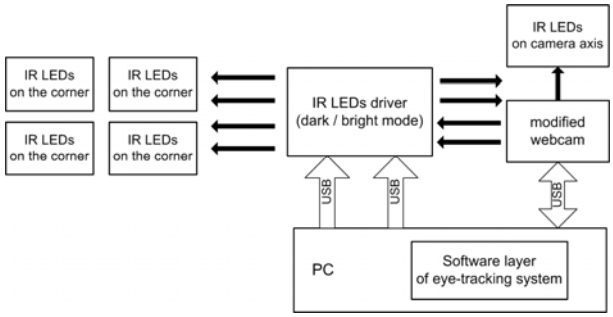

Figure 2. Block diagram of the developed gaze-tracking system [4].

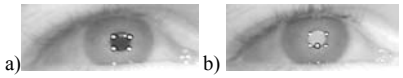

Figure 3. Part of the image with the eye region detected. Eye is illuminated by the IR LEDs: (a) in the dark eye mode, and (b) in the bright eye mode.

B. Gaze direction estimation algorithm

Four modules of IR LEDs placed in display's corners produce unique corneal reflections (glints). The algorithm analyses each video frame taken by camera watching the user's face and locate glints. Relations between the IR LEDs on the display and reflections on the eye are shown in Fig. 4. Known coordinates of characteristic points (the pupil centre and four glints) are sufficient to determine the fixation point.

To locate glints in the image, the frame is segmented for the processing into sub regions of 40x40 pixels. First, the brightest points are determined in each region. When the configuration of the brightest points in a segment form a quadrangle shape with edges and angles fulfilling certain conditions, then this segment is considered as containing the first eye with visible glints. The area for detection of second eye is limited to segments located on the left and right side of the first candidate segment. Precise coordinates of each glint in both eyes are then determined. It is performed by calculation of centers of gravity (COG) of areas representing glints (consistent areas of high intensity pixels). Next, the region of pupil is detected in the eye image through ellipse fitting. Coordinates of the center of ellipse (approximating the shape of the pupil) are taken as coordinates of the pupil center. Then, the fixation point is estimated for each eye independently, based on the location of all detected points: glints and pupils. The last step consist of correction of the fixation point coordinates, based on data obtained in the calibration procedure, for particular user and light conditions. Functional scheme of the algorithm is presented in Fig. 5.

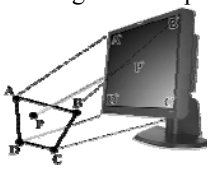

Figure 4. Relations between glints and IR LEDs on the display.

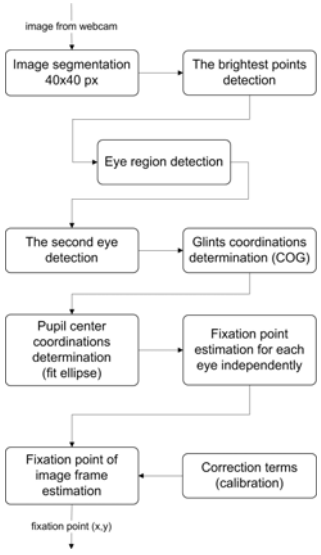

Figure 5. Block diagram of fixation point estimation algorithm.

Spatial resolution tests of the described gaze-tracking system were conducted. Five persons took part. The users were sitting 70cm from the 22'' LCD computer screen with resolution of 1024x768. They were instructed to look for 3 seconds at any point of the screen, then point that location with mouse cursor. If estimated gaze direction matched cursor location (with tolerance of 50 pixels vertically and horizontally), then the result is taken as correct. Averaged result spatial resolution is not consistent throughout the screen. It was noticed that in the middle and upper parts of the monitor screen the resolution is sufficient to distinguish 70 areas (7 rows x 10 columns), but in the lower part of the screen only 14 regions (2 rows x 7 columns). Therefore, it can be assumed that the system's resolution is approximately 8 x 7 regions. Thus, the effective gaze control of 9 monitoring cameras simultaneously is ensured.

C. PTZ camera steering

PTZ camera steering presented in this paper is based on user's gaze direction. The user is able to change by gaze a field of view of the PTZ camera. Preview of nine PTZ cameras is presented on the screen. Video sub window of the selected camera is enlarged when the user looks at it for 3 seconds. Then the control mode of PTZ module of the camera is activated. The PTZ camera is moved when user is looking at the vicinity of the video frame border, e.g. if the system detects the user's gaze fixation point near one of the edges or corners of the frame, then a control signal is sent to the camera, rotating it in selected direction as long as the user focuses on a particular point in frame boundary. While gazing at the center of the frame the camera remains still. The PTZ mode switch back to 9 video previews, when the gaze point leaves the frame area for more than 2 second.

The application of gaze-tracking system responsible for communication with cameras has been implemented in C# programming language. The presented application is easily configurable by editing of XML files containing cameras configuration, up to 56 IP camera addresses, and their locations on computer screen grid. Fig. 6 presents the system during operation.

The described camera control procedure can be utilized for teleconference systems, where each side have gaze-tracking systems, and steer far-end camera to look at particular interlocutor.

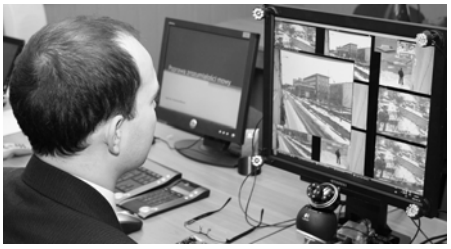

Figure 6. The user steering PTZ camera with gaze-tracking system. The first video stream in middle row is enlarged with camera steering mode engaged.

D. Usability study

Users experience with the application was studied. Assessed parameters gave an overall view on the effectiveness of PTZ cameras steering. Five Ph. D. students (25-27 years old) participated in the study. They did not have prior experience with operating monitoring systems. The test for each person lasted 10 minutes. During that time, the respondents were repeatedly choosing camera views by gaze. After enlarging the preview, they were instructed to freely steer the camera (simulation of real working conditions). All assessment parameters were evaluated on a 1-5 scale (1 - poor/low, 5 - very good/high) (Table I). The last column of the table contains the mean value from all respondents (*Mean*). The first parameter was the *comfort of working with the application*. Next was the opinion on the *usefulness of the application* in use for the monitoring system. The last was the overall *subjective assessment of application*. Based on the results, it was concluded that the usage comfort is the worst feature of the application - the mean value of all respondents was 3,6, but respondents considered that the technology can be functional in this type of applications (4,2) and the general assessment was relatively high (4).

TABLE I. TEST RESULTS

Parameter	**The tested person**					**Mean**
	1	***2***	***3***	***4***	***5***	
comfort of using	3	4	3	4	4	3,6
usability of application	5	4	4	3	5	4,2
general assessment	4	5	4	3	4	4

The evaluation of usage comfort indicate that the application in its current form is not yet ready for widespread use in CCTV systems. Nevertheless, the fact that the system is hands-free is important, because the physically disabled persons and others with problems with using computer mouse have the opportunity to use such systems. Comfort of using this application for such persons would certainly be higher than the rating of Ph. D. students. The next stage of research on the usability of the application will involve physically disabled people.

III. ACOUSTIC SIGNALS PROCESSING

A concept, a practical realization and applications of a passive acoustic radar (PAR) to automatic localization and tracking of sound sources are presented below. Contrary to active radars, PAR does not emit a scanning beam, but after receiving surroundings sounds it provides information about the directions of sources. The device consist of a new kind of multichannel acoustic vector sensors (AVS) [6] and dedicated digital signal processing algorithms developed by the authors. Concerning the acoustic properties, the classical beamforming arrays have limited frequency range and a line (or plane) symmetry with significant decreases of resolution at directions far from symmetry axis. The AVS approach is broad-banded, works in 3D, and has a better robustness [7].

The ability of a single AVS to rapidly determine the bearing of a wideband acoustic source is of essence for passive monitoring systems.

In the PTZ gaze-tracking system the audio processing is utilized in two modes: "audio slave" and "audio master". While the user operates the PTZ camera by gaze tracking, the PAR algorithm is in the "audio slave" mode, performing adaptive changes of sound directivity characteristics 45° wide, respective to camera direction. That allows presenting only the sounds incoming from the view direction. In case the PAR algorithm detects an important sound (for more details see [8][9]), the system is switched into the "audio master" mode, and the camera is automatically steered to the direction of the sound source, and operator is informed of detected sound event. Then gaze steering is again switched on, therefore the user can further investigate the scene.

Foundations of audio processing method and details of both modes are presented below.

A. 3D sound intensity probe (Acoustic Vector Sensor)

The single AVS, measures the air particle velocity instead of the acoustic pressure [10]. The air velocity is measured in the direction perpendicular to two tiny resistive strips of platinum that are heated to about 200°C (Fig. 7). The air flow cools down both sensors, and alters their resistances. Therefore one can measure air speed and distinguish between positive and negative velocity direction. It is much more sensitive than a single hot wire anemometer and is (almost) not temperature sensitive [11]. Velocities in the range of 10 nm/s up to 1 m/s can be measured.

Three orthogonally placed particle velocity sensors have to be used to determine azimuth and elevation of the direction vector. With addition of a pressure microphone, the sound field in a single point is fully characterized and also the acoustic intensity vector, which is the product of pressure and particle velocity, can be determined [12]. This intensity vector indicates the acoustic energy flow. The full three dimensional sound intensity vector can be determined within the full audible frequency range 20 Hz up to 20 kHz.

B. Sound source direction detection – "audio master"

The algorithm of the PAR is based on 3D sound intensity component (Fig. 8). First, the acoustic signals are captured. Then the dominant frequency of the sound is estimated based on the FFT coefficients and using Quinn's First Estimator [13]. Next, the frequency value is used to design the narrow-band recursive filter [14]. The result of the filtration is finally used to compute the particular sound intensity components.

The time averaged intensity I in a single direction is given by (1) [15]:

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

where: $p(t)$ – sound pressure (scalar),
$u(t)$ – particle velocity (vector).

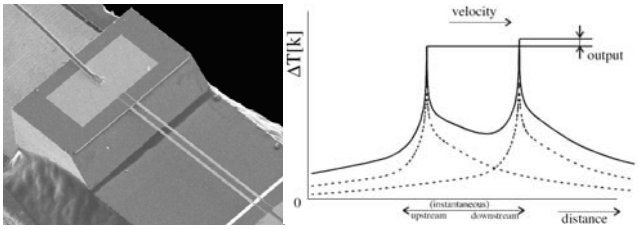

Figure 7. Microscope picture of a standard Microflown probe (left). Temperature distributions (right): dotted line: temperature distribution due to convection for two heaters - both heaters have the same temperature. Solid line: sum of two single temperature functions: a temperature difference occurs [11].

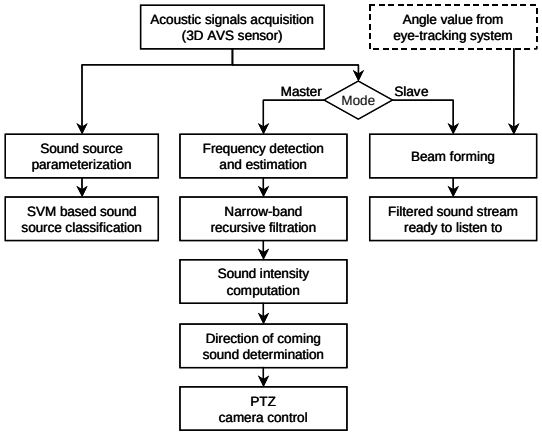

Figure 8. The block diagram of the PAR algorithm.

In the presented results the time average T (1) was 4096 samples (sampling frequency 48kHz), therefore the sound direction was updated more than 10 times per second.

Tests of sensitivity and accuracy of PAR were conducted in anechoic chamber and in typical reverberant conditions. Pure tones and 1/3 octave band noises from 125 to 16000Hz were used. The AVS and reference measurement microphone (Bruel&Kjær PULSE system type 7540 with microphone type 4189) were located in the same place to ensure identical acoustic condition. The sensitivity of PAR was expressed by the SPL value, and its accuracy in the noisy conditions was expressed by the sound to noise ratio (SNR) (3).

$$SNR_{dB} = SPL_{Signal\ dB} - SPL_{Noise\ dB} \qquad (3)$$

The obtained sensitivity of PAR in the anechoic chamber was 45dB. The SNR_{dB} results for both kinds of test signals are presented in Fig. 9. The noise source was 62dB. The expected angular resolutions were set at: ±1°, ±3°, ±5°, ±10°, ±15°, ±30° and ±45°. PAR performs well in continuous localization for all considered signals. The results of the sound source localization in real environments are presented in Fig. 10a. An alarm signal revolver was used as a sound source. For every location three shots were generated. The averaged positions are shown. Mean Squared Error (MSE) for the *angle* and (x, y) *coordinates* estimation were calculated. The *angle* estimation MSE was 5.5°, and the *coordinates* MSE was 0.7m.

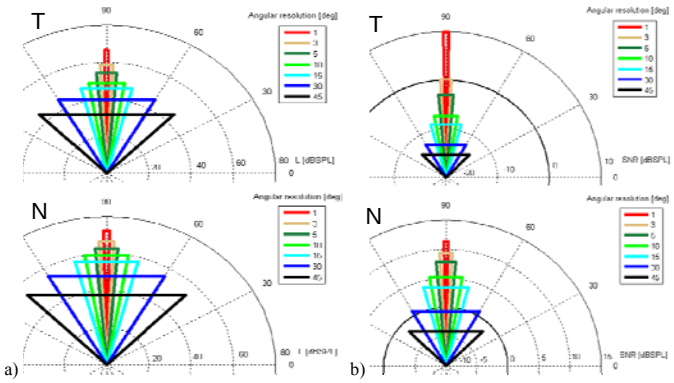

Figure 9. Angular resolution of sound source localization for various source levels L and various noise conditions SNR. Triangle widths portray resolutions, triangle heights match L and SNR values on radial plot. Averaged results for all examined frequencies: a) absolute sensitivity L [dBSPL], b) averaged SNRdB results. "T" – pure tones, the recursive filtration was applied, "N" – noisy test signals.

C. *Sound source direction filtration – "audio slave"*

If the sources are in the far field, it is possible to create a virtual microphone with a variable directivity pattern. The pressure microphone has an omnidirectional pattern and the Microflown probe has a figure of eight pattern. The axis of zero sensitivity (ZS) varies when a summation is made with the sound pressure signal (p) and particle velocity signal (u). The ZS is at 90° for a pure velocity signal u, at 0° for $p+u$, and at 180° for $p-u$. So ZS is "steered" by the ratio of p and u in the summation (Fig. 10a). ZS is sharp therefore one can find out if there are one, two or more sound sources [11].

For further modification of this virtual microphone a summation or subtraction of signals p and u can be performed. A cardioid or unidirectional types of directivity can be achieved (Fig. 10b, c).

IV. Results

The presented system was used to effectively control 9 PTZ cameras (the number limited by network capability for video streaming). The users reported fast habituation and intuitiveness of hands-free operating of the cameras.

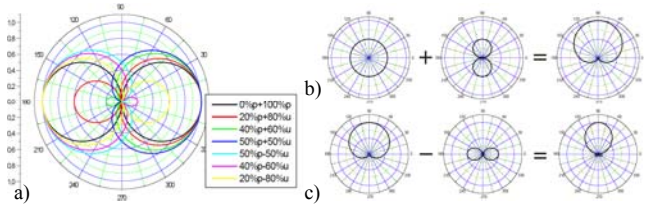

Figure 10. Directivity patterns obtained by combination of a microphone and probe signals: (a) steering of line of zero sensitivity by pressure signal p and velocity signal u summations, (b) omnidirectional microphone characteristic (left) summed with a figure of eight normal probe characteristic (middle) creates a cardioid microphone characteristic (right), (c) The response of a cardioid (left) minus the squared response of the lateral velocity probe (middle) results in a response that is almost similar to an unidirectional microphone [11].

In the future the system will be tested for 56 streams, utilizing full resolution of gaze-tracking system. The system can be easily introduced in monitoring with PTZ cameras or any application requiring hands-free camera control.

A concept and testing results of the passive acoustic radar (PAR) were presented in the paper. Diverse test signals were used. Based on the results the small value of signal to noise ratio is sufficient to localize sound source effectively (SNR_{dB} near to 0 dB). The application of the recursive filtration can improve sensitivity and accuracy (SNR_{dB} below -10 dB for tonal components). The filtration can be used to discriminate between multiple sources. PAR algorithm properly detects and localizes the source in real acoustic conditions. The automatic and continuous tracking of the sound source movements in real time was also possible. Sound classification can be used to extend system usefulness, and improve the functionality of traditional surveillance.

Acknowledgment

Research is subsidized by the Polish Ministry of Science and Higher Education within Grant No. R00 O0005/3 and by the European Commission within FP7 project "INDECT" (Grant Agreement No. 218086).

References

[1] Z. Zhu, Q. Ji, "Eye gaze tracking under natural head movements", Proc. of the 2005 IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, 2005, pp. 918-923.

[2] B. Kunka, B. Kostek, "A new method of audio-visual correlation analysis", Proc. Int. Multiconference on Comp. Science and Inform. Technology (IMCSIT), vol. 4, 2009, pp. 497 - 502.

[3] B. Kunka, B. Kostek, M. Kulesza, P. Szczuko, A. Czyzewski, "Gaze-Tracking-Based Audio-Visual Correlation Analysis Employing Quality of Experience Methodology", Intelligent Decision Technologies, IOS Press, in press.

[4] B. Kunka, B. Kostek, „Exploiting Audio-Visual Correlation by Means of Gaze Tracking", International Journal of Computer Science and Applications, 2010, in press.

[5] "Safety of laser products, Part 12: Safety of free space optical communication systems used for transmission of information," IEC 60825-12:2004, CENELEC 2004.

[6] Microflown Technologies – Home: http://www.microflown.com/

[7] M. Hawkes, A. Nehorai, "Wideband Source Localization Using a Distributed Acoustic Vector-Sensor Array", IEEE Trans. on Signal Processing, vol. 51, no. 6, 2003.

[8] P. Zwan, A. Czyzewski, "Automatic sound recognition for security purposes," Proc. 124th Audio Engineering Society Convention, Amsterdam, 2008.

[9] P. Zwan, P. Sobala, P. Szczuko, A. Czyzewski, "Audio Content Analysis In the Urban Area Telemonitoring System," Multimedia Services In Intelligent Environments, 2009.

[10] H.E. de Bree The Microflown: "An acoustic particle velocity sensor," Acoustics Australia 31, 2003, pp. 91-94.

[11] H.E. de Bree, "The Microflown," E-book: http://www.microflown.com/r&d_books_Ebook_Microflown.htm

[12] J. de Vries, H.E. de Bree, "Scan & Listen: a simple and fast method to find sources", SAE Brazil (2008)

[13] M. Donadio, "How to interpolate the peak location of a DFT or FFT if the frequency of interest is between bins," http://www.dspguru.com/dsp/howtos/how-to-interpolate-fft-peak.htm

[14] S.W. Smith, "The Scientist and Engineer's Guide to Digital Signal Processing", California Technical Publishing, 1997.

[15] T. Basten, H.E. de Bree, E. Tijs, "Localization and tracking of aircraft with ground based 3D sound probes", ERF33, Kazan, Russia, 2007.