Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect)   Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D6.6
First audio-visual data collection with AWEAR and OHSU system

Date of deliverable:31.03.2009
Actual submission date: 08.05.2009

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **Fraunhofer Institute Digital Medientechnologie, Project group Hearing, Speech and Audio Technologies**

Revision: [1]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

DIRAC

Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

D6.6 FIRST AUDIO-VISUAL DATA COLLECTION WITH AWEAR AND OHSU SYSTEM

Fraunhofer Institut Digitale Medientechnologie, Project Group Hearing, Speech and Audio Technology (FRA)
Oregon Health & Science University (OHSU)

Abstract:

One of the objectives of WP6 is
“Recording audio, visual, and audio-visual databases that would support DIRAC research thrusts”.
For this purpose the partners have defined two application domains: the security and surveillance on the one hand, and in-home monitoring of elderly people on the other. For both application domains scenarios have been developed.
In both domains there is a need for the DIRAC methods to be applied, since unobtrusive automatic monitoring is required, which will reliably react to unforeseen circumstances and events.
In both domains the current standard is the use of fixed cameras, but also in both domains there are developments towards the use of sound information and the use of mobile platforms. For this reason audio-visual recordings are planned with both a fixed and a mobile platform. The mobile platform is a new and much improved version of the AWEAR I platform used previously in the DIRAC project. First recordings with the new AWEAR II platform have already been carried out.
Recordings for the security scenario will be carried out, using the AWEAR II platform, by FRA. Recordings for the in-home care scenario will be carried out, with identical hard- and software but in platform, by OHSU in their Point of Care Laboratory.

Table of Contents

1. Introduction

In the new TA, submitted at 12th of February 2009, application scenarios were defined for the DIRAC technology. Based on the experience and background of dr. ir. van Hengel, of the new partner FRA, in the security market and the high demand in this market for automated and intelligent surveillance systems, it was concluded that this would be one of the domains that would benefit considerably from the technology developed in the DIRAC project. The focus of FRA and OHSU on applications in in-home monitoring of elderly people made this market a logical second application domain.

In both domains there is a need for 24/7, unobtrusive, autonomous, and therefore intelligent, monitoring systems to assist human observers. The use of sound to augment camera surveillance has recently been introduced in the security market (van Hengel & Andringa, 2007) and is planned in in-home monitoring (van Hengel & Anemüller, 2009). Spotting and properly responding to unforeseen situations and events is one of the crucial aspects of monitoring systems in both application domains.

For both application domains scenarios have been developed that will show the potential of the DIRAC theoretical framework and the techniques developed in the various workpackages, while attempting to address realistic and interesting situations that can not be handled properly by existing technology. For the security domain audio-visual recordings will be made primarily of outdoor scenes, where the use of the AWEAR II mobile recording device provides the required flexibility. These recordings will be made by FRA. For the in-home care domain recordings will be made in the Point of Care Laboratory at OHSU, which provides realistic conditions as well as in-depth insight into the needs and requirements of this market segment.

2. AWEAR II hardware

Based on experiences gained with the AWEAR I it was decided to design a new version of the AWEAR recording platform. The main objective was to develop a more portable and wearable platform and ensure synchronization of audio and video recording, while retaining high resolution of both audio and video.

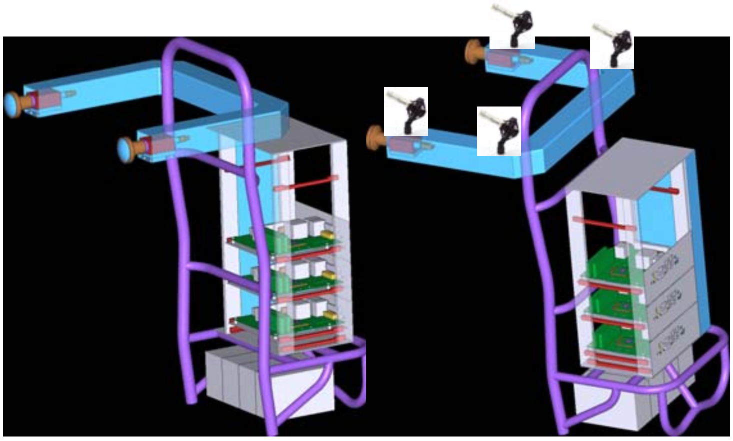

Figure 1: AWEAR II design

The AWEAR II design was to be used as a recording platform first and foremost, but the option that some (pre-)processing of audio and video material could take place on the device itself was taken into account during the design phase.
In order to be able to use the AWEAR II for outdoor recordings and allow flexibility, the system had to be wearable and self-powered. It should be able to function without the need for recharging for several hours.

The components chosen were:

Video:

AVT Stingray camera	2
Fujinon FE 185CO86HA1 fish-eye lens	2

Audio:

T-bone EM700 stereo mic set	2 (4 mics in total)
FOCUSRITE Saffire PRO 10	1
Sennheiser EH 350 headphones	1

PCs:

Siemens D2703-S mini ITX boards	3 (2 video, one audio & control)

Frame:

Tatonka Lastenkraxe backpack	1
Camden 12V gel batteries	4

A multitude of custom parts, including the triggering system, were developed by KUL, who also handled construction and hardware-testing of the AWEAR II. Weighing a total of ~20 kgs the platform can easily be carried around and can be autonomous for ~3 hrs.

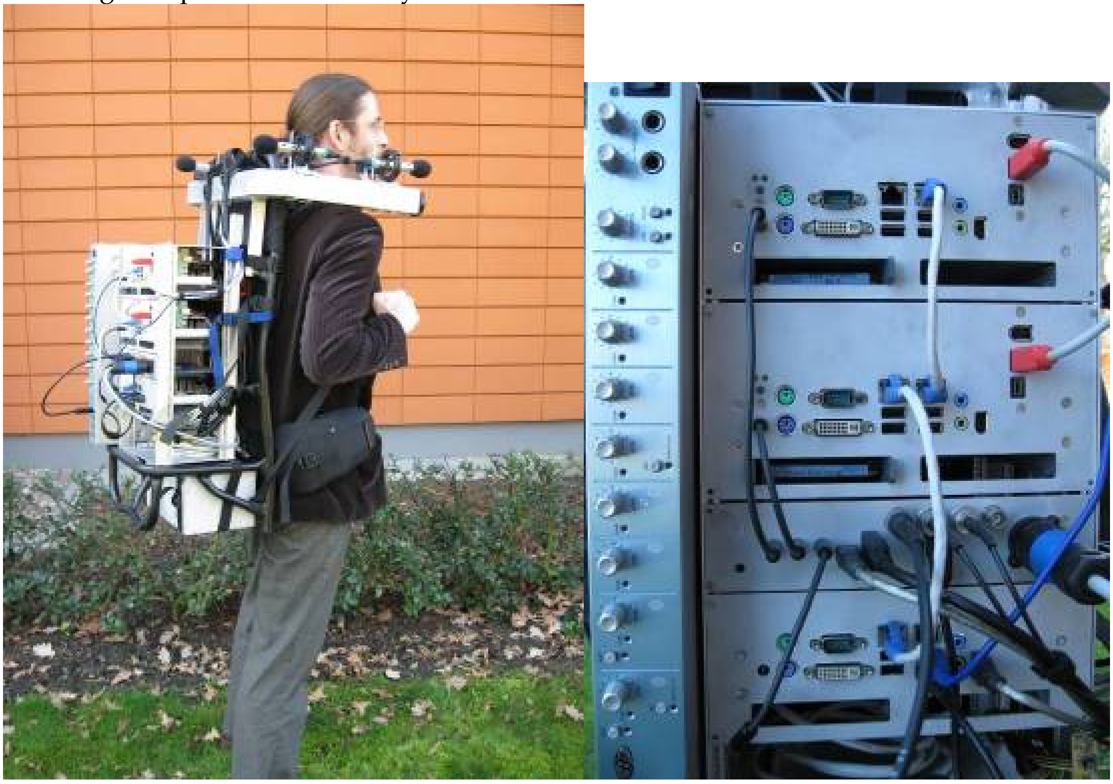

Figure 2: The AWEAR II hardware

3. AWEAR II software

The custom made triggering hardware provided by KUL signals the cameras to capture images. These images are saved, each on a separate Siemens mini-computer, by custom made Violin software, provided by KUL. The images are saved in raw format into dump files containing 1000 images each.

The audio is captured from the Focusrite connected by Firewire onto the third mini-computer using ffado and jack_capture open source software. Up to 8 channels of audio can be recorded, of which 4 are used for the 4 microphones and 1 is used to record the triggering signal. This audio channel is used in the off-line cutting of the scenes for synchronization of the audio and video material.

All three mini-computers are connected via LAN, and via a wireless USB to a portable ASUS EE PC. Custom made software on this EE PC is used to initialize the system - e.g. adjust camera settings to lighting conditions -, start and stop recordings.

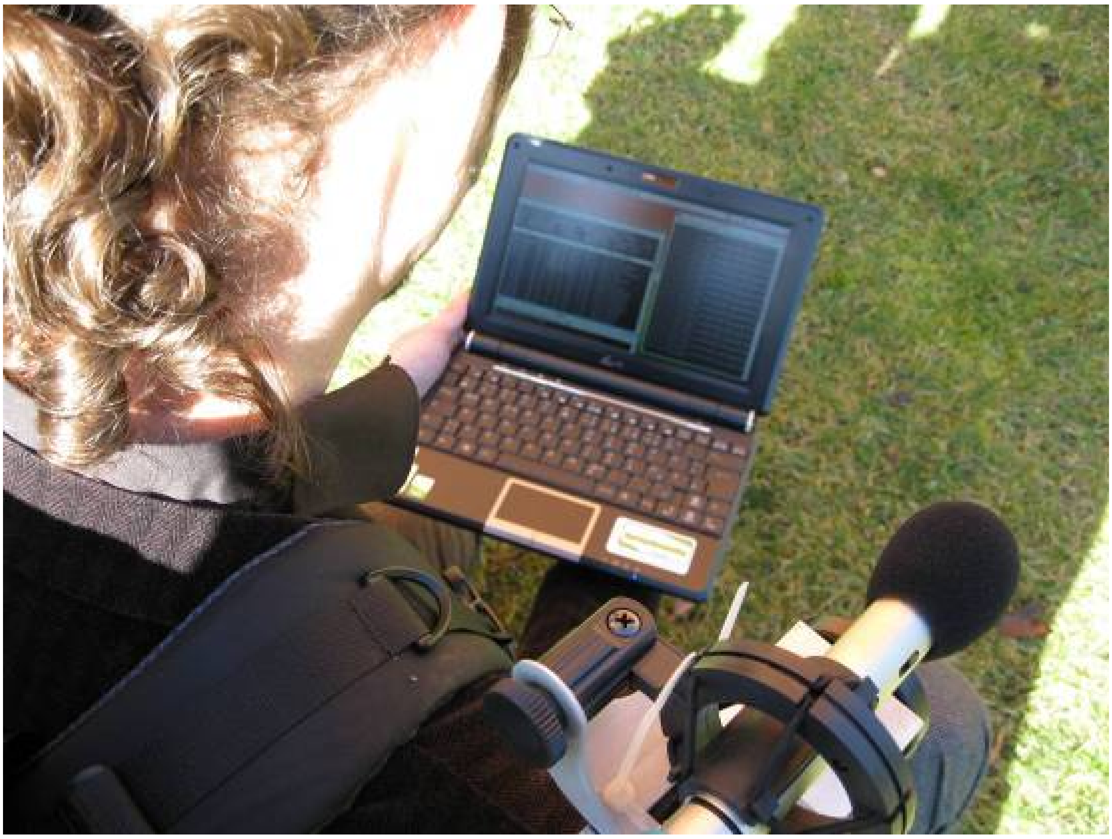

Figure 3: AWEAR user with EE PC for control of recording

Once a recording session has finished the three hard disks, containing the video signals from the left and right camera and the 8 track audio respectively, are removed from the AWEAR II and connected to a dedicated PC via ESATA HDD cradles. The proper parts of the recordings are selected manually and, using custom developed software, the corresponding images from both cameras and the 4-channel audio signal are stored. The images and audio are transported to the DIRAC-server, where they can be accessed by all partners. Further processing of video and audio signals is performed at CTU, ETHZ and OL, as described in D1.7.

4. AWEAR II Recording session I

Oldenburg, 24-26 November 2008
The application scenario for the security market consists of a situation with multiple pedestrians and vehicles, all requiring more or less attention and creating incongruent and potentially dangerous situations, such as collisions.
Based on the developing description of this application scenario initial recordings were made at the car park of the House of Hearing in Oldenburg.

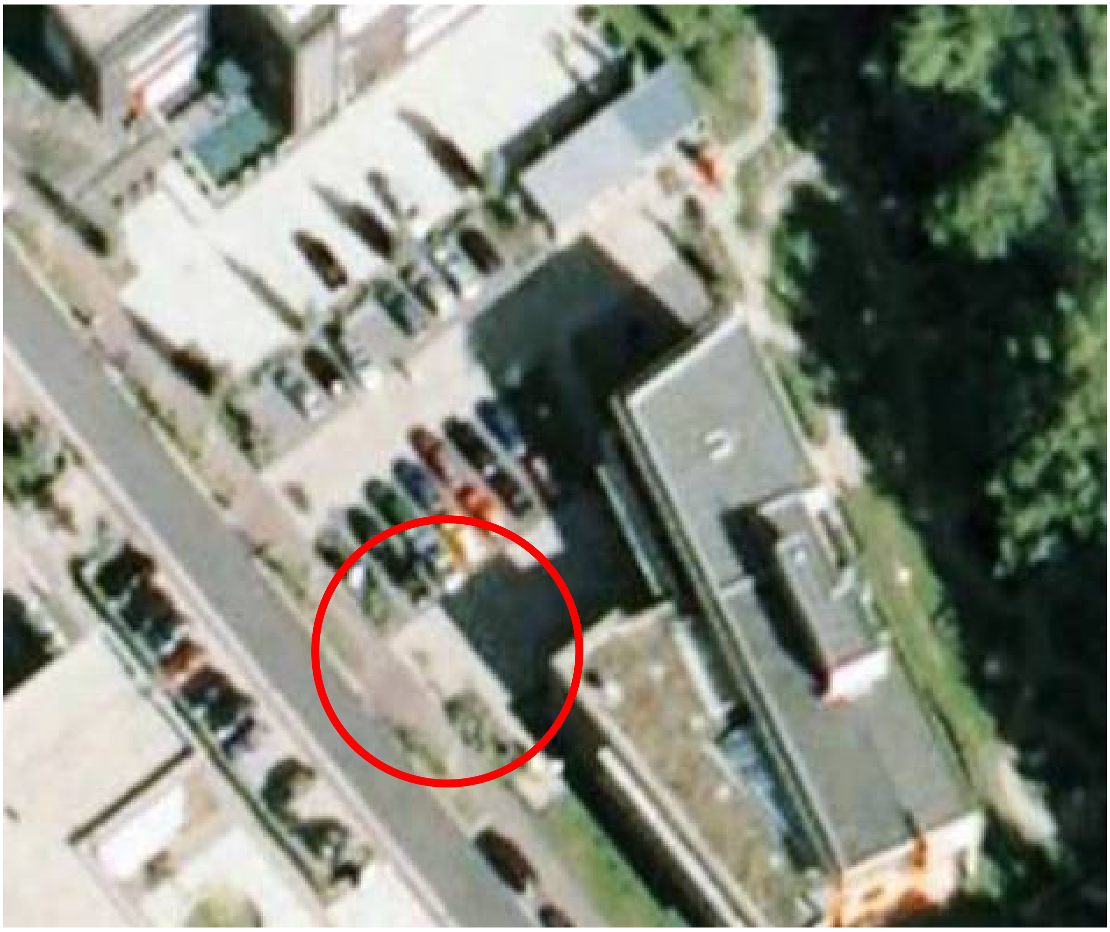

Figure 4: car park of the House of Hearing

A person wearing the AWEAR walks along footpath. A person on a bicycle approaches him from behind. The person on the bicycle overtakes the person with the AWEAR, while several pedestrians walk along the footpath and across the car park and cross the street.

Several recordings were made, in some of which the person on the bicycle shouts and swears as he approaches, in some a pedestrian crosses just as the bicycle passes and there is almost a collision.

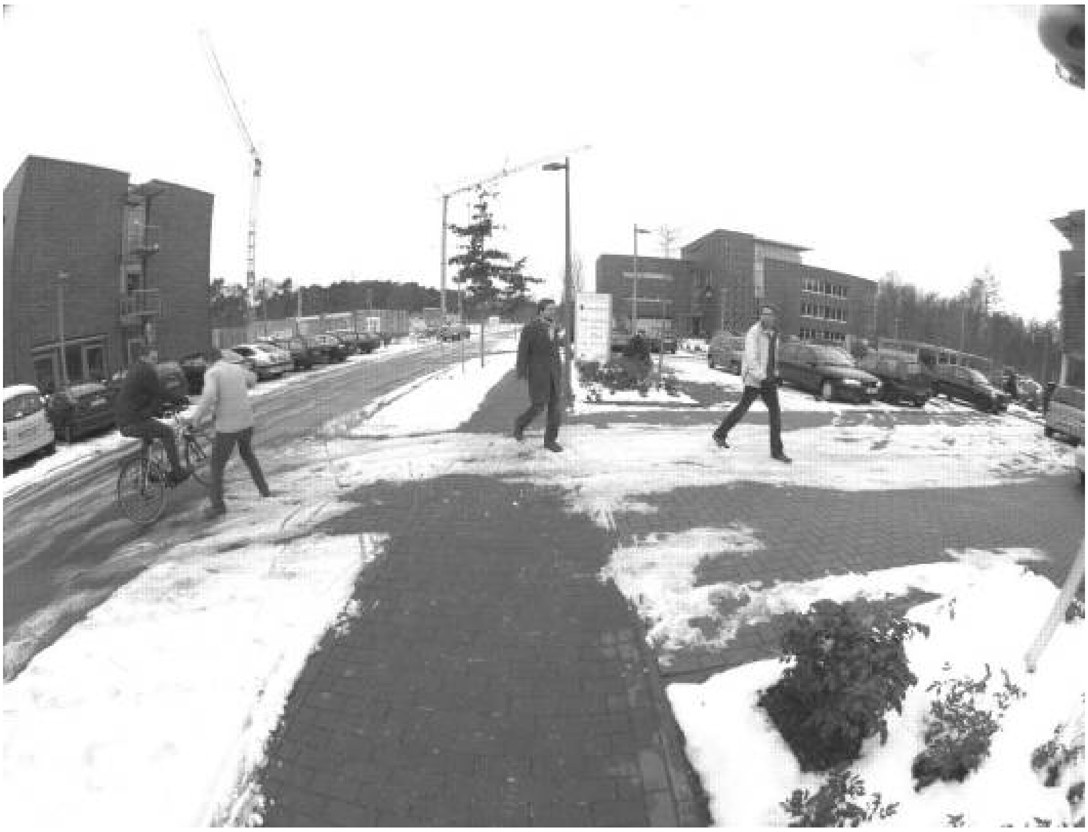

Figure 5: Image taken from the recording session with almost collision between bicycle and pedestrian

One of the strengths of the DIRAC consortium is the fact that it integrates work in audio and in video. It is evident that the scenarios need to reflect this integration. Also, there is increasing interest in multi-modal information fusion, making this an excellent topic to demonstrate the DIRAC framework on.

As a simple first step towards a demonstration of audio-video fusion a simple scenario was devised purely as a demonstration of the DIRAC theoretical framework in an audio-video integration task.

This demonstration is described in more detail in D1.7, and consists of a triplet of detectors for speaking persons. One detector is trained only on audio material, one is trained only on video material, and one is trained on audio and video. A logical combination of the output of the audio detector and the output of the video detector should equal the output of the detector that uses audio and video as input. That is, if everything is as expected, or as seen in training.

In order to show the effect of the DIRAC theory the three detectors are confronted with recordings of a person standing next to a loudspeaker, with speech coming from the loudspeaker only. The idea is that the detector trained on the audio and video features has implicitly been fed information on the spatial distribution of the audio and video features and has 'learned' that the two need to be collocated. Therefore this detector will not give a positive response on the scene with a silent person and a 'speaking' loudspeaker. The logical combination of the outputs of the separate detectors for audio and video does not take the collocation requirement into account and will give a positive output on the scene where there is speech – coming from the loudspeaker – and there is a person in view – not speaking-.

Several recordings were made with a person and a loudspeaker mounted on a tripod in view of the AWEAR. In some the person was speaking, in some the person was quiet. In some the loudspeaker was silent in some it was producing speech.
Also some recordings were made of a person walking and of more than one person in the field of view.
All recordings were made inside the KAS in the House of Hearing in Oldenburg. The KAS – Kommunikations Akustik Simulator – is a specially designed room measuring approximately 12 meters by 7 meters. The walls, ceiling and floor of the room are specially treated with acoustically absorbing materials and so-called Schröder diffusers are mounted on the walls to minimize sound reflections. This gives the room unnaturally 'dry' acoustics. Furthermore, 16 microphones are distributed along the ceiling and 12 loudspeakers. An additional 12 flat panel loudspeakers are located on the walls. The sound signals received by the microphones are fed to a matrix processor which artificially introduces delay and gain and plays back the signals over the loudspeakers, thus artificially introducing reflections. By modifying the strength and delay of these reflections, as well as their spectral characteristics, the acoustics of the room can be altered to mimic a room smaller or much larger than the actual dimensions. In addition to altering the perceived dimensions of the room, the reflective qualities of the room can also be altered to range from 'acoustic heaven' to 'acoustic hell'.

5. AWEAR II Recording session II

Oldenburg, 19 January 2009.
For more thorough development and study of the short term demonstrator for AV interaction, described in the previous section and in more detail in D1.7, additional recordings were required. All recordings were made in the KAS.
This time the recordings were made with the person at positions indicated in the following figure:

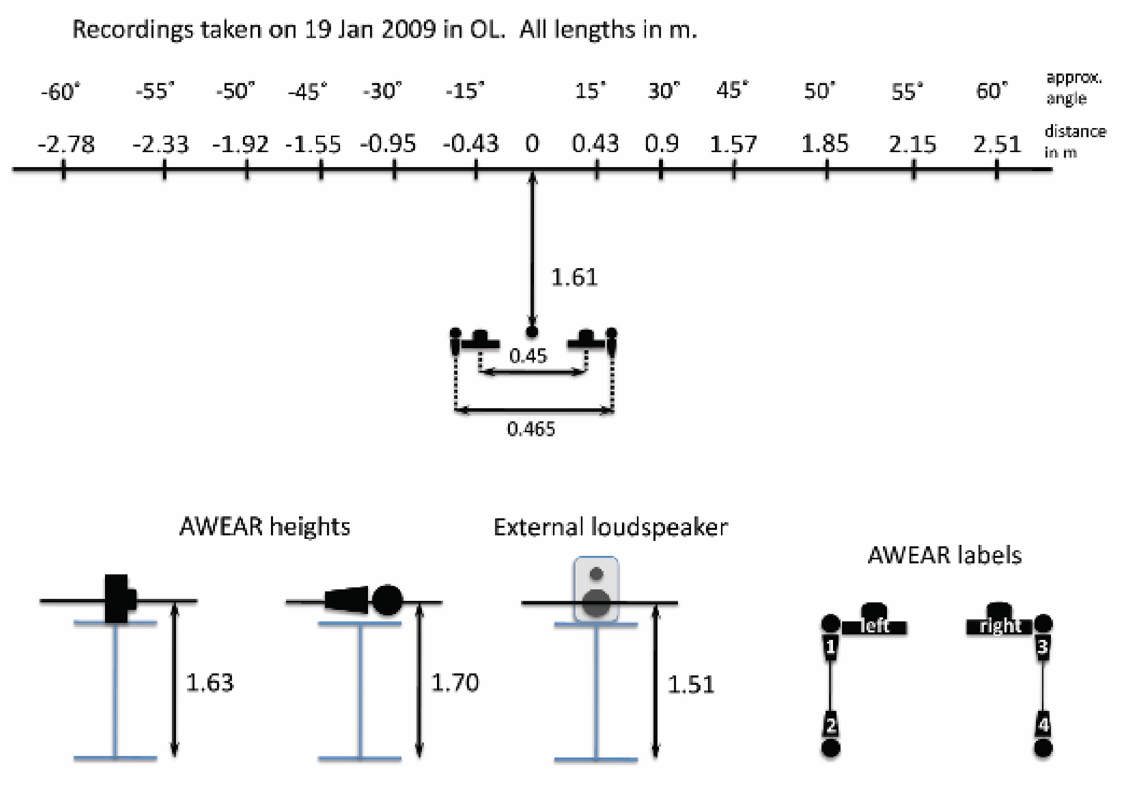

Figure 6: recording setup for AV demonstrator recordings

Calibration recordings were made with the loudspeaker at each of the 13 positions. Recordings were made with a person and loudspeaker where the loudspeaker would produce speech and the person was silent for ~2 minutes. Then the person would also start talking for ~2 minutes and then the loudspeaker would be turned off and only the person would be speaking for an additional ~2 minutes.

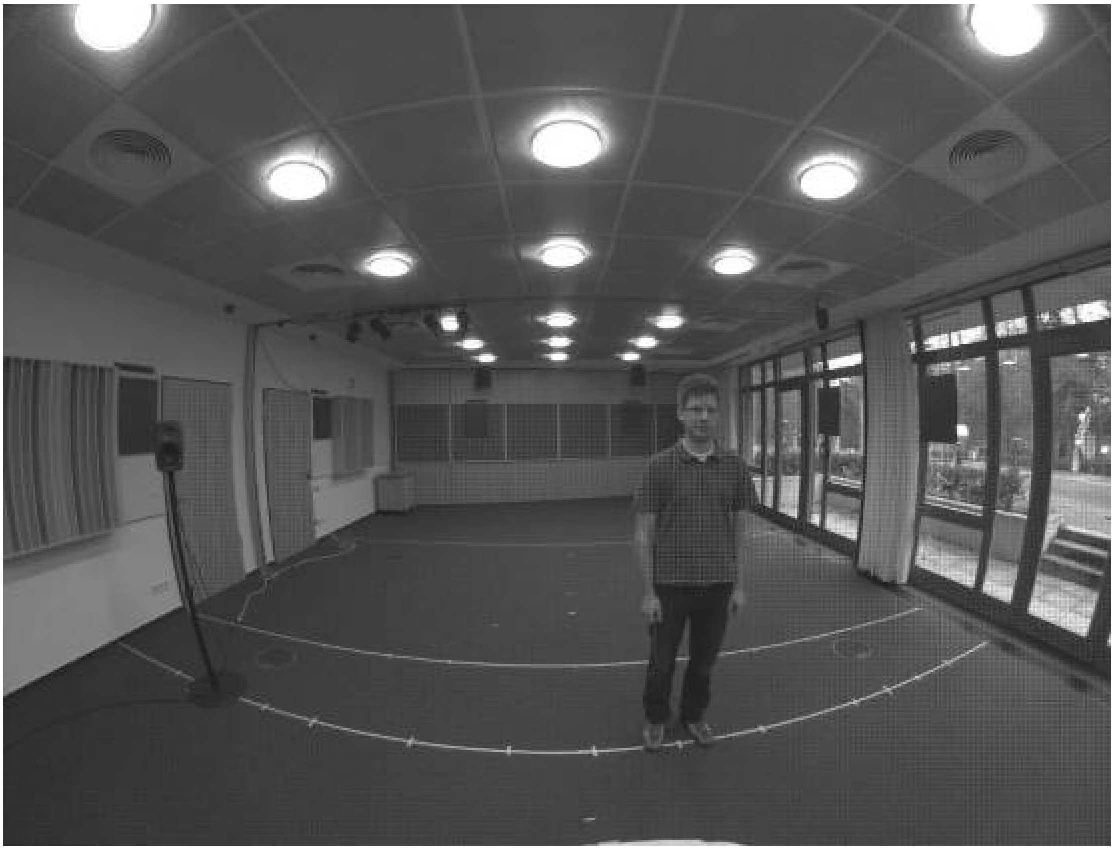

Figure 7: image from the recording of the AV demonstrator training data in the KAS

Additional recordings were made with a person walking along the line connecting the positions. The person would either be talking or silent and would either be facing the AWEAR or facing front, leading to four combinations.
Five persons were recorded.

6. AWEAR II Recording session III

Oldenburg, 5 February 2009.
As an additional demonstrator of the DIRAC framework and the status of development at ETHZ, it was decided to show the multi-layer tracking on a pedestrian. For this case pedestrians would be walking in a straight line and suddenly change their motion pattern.
Three deviations from normal walking style were used:
- sudden stop, here the person would pretend to startle and suddenly stop
- falling, here the person would pretend to trip and almost fall
- running, here the person would pretend to be frightened and start running

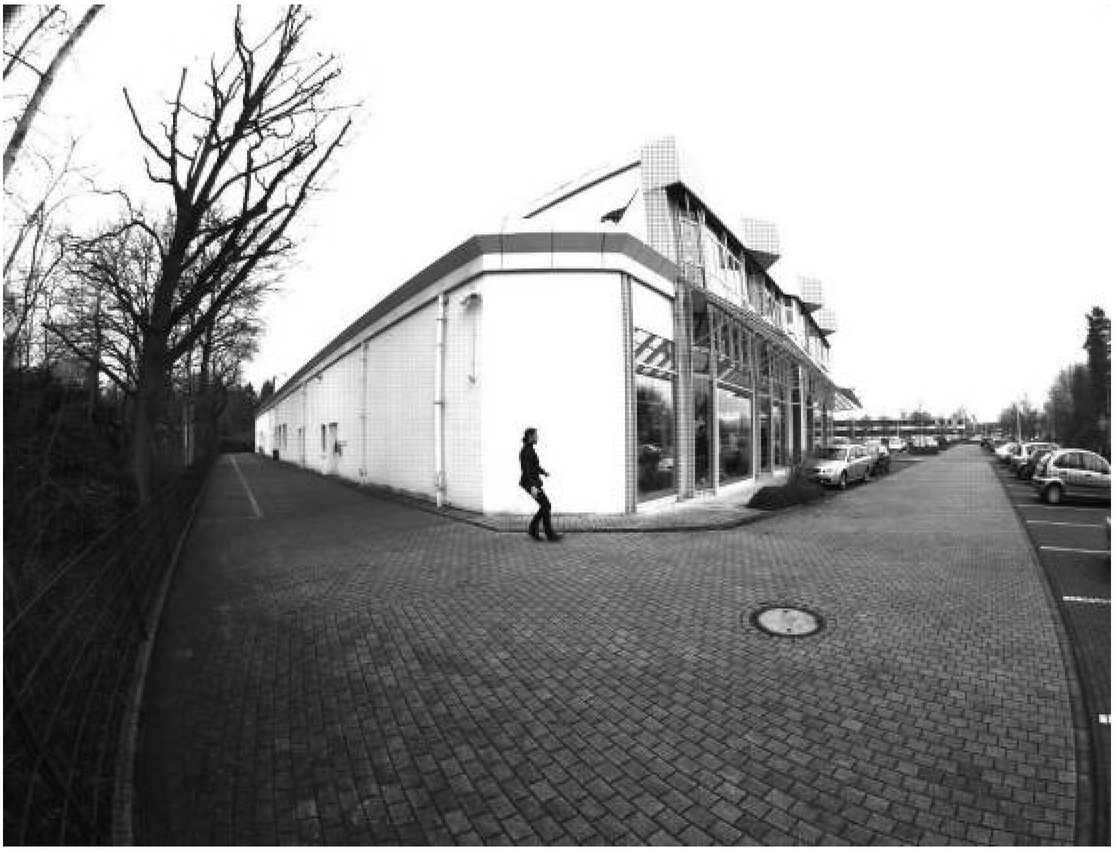

Figure 8: image taken from the recording session at Bruno Kleine, showing a pedestrian suddenly stopping

The recordings were made at the car park of the Bruno Kleine store ~500 meters from the House of Hearing. This location was chosen because of a white wall which provided a simple background.

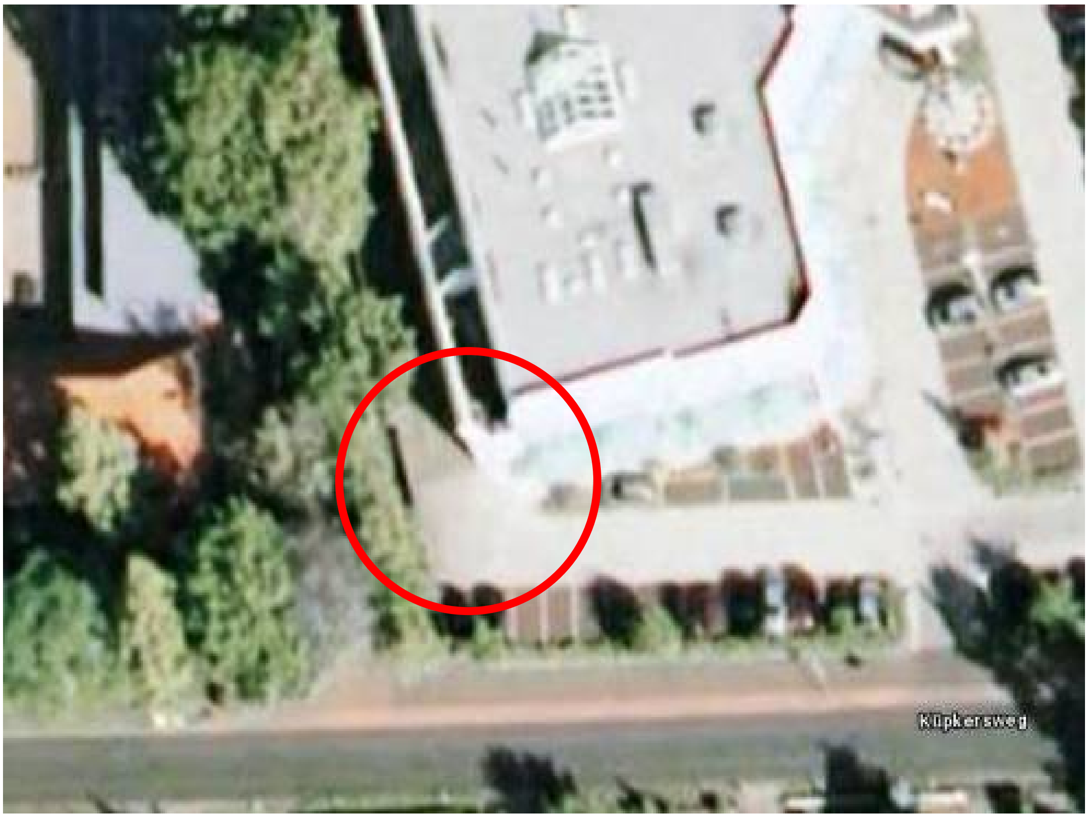

Figure 9: the Bruno Kleine car park indicating where the recordings were made

Five persons were recorded.

7. AWEAR II Recording session IV

Oldenburg, 3-7 April 2009.
During the review meeting in Oldenburg several recordings were made to show the operation of the AWEAR II platform and for use in the short term demonstrators. These recordings include a recording for the AV demonstrator with a person walking into the field of view while a loudspeaker in the left of the field of view is producing speech. After ~1 minute the loudspeaker stops and the person starts speaking. While speaking the person starts walking again and walks out of the field of view.
Another scene was recorded outside with a number of pedestrians, one of which approaches the AWEAR user and starts a conversation. A bicycle passes and suddenly turns towards the AWEAR user, while some people shout.
As a first attempt at providing data for vehicle detection recordings were made of passing cars and bicycles.
Also recordings were made with a walking AWEAR user who would diffuse the images from one of the cameras by taping a piece of foil in front of it. These recordings will be

used by CTU to check whether the DIRAC framework could be used to detect low-level camera issues and take appropriate action (see also D1.7 for more detailed description). Finally, recordings were made of the AV scenario with a moving AWEAR. So far, all recordings had been made with a stationary AWEAR II placed on a table. This was mainly done to avoid having to carry the AWEAR for long periods of time. In order to show the full preprocessing chain for this scenario, the recordings were repeated with the AWEAR worn by a user.

8. OHSU hardware

In the Point of Care Laboratory (PoCL) the experiments were designed to capture all activity within the living room and kitchen of a model apartment of an elderly person. The PoCL is a fully functioning apartment where scenarios can be acted out using actors or actual elderly.

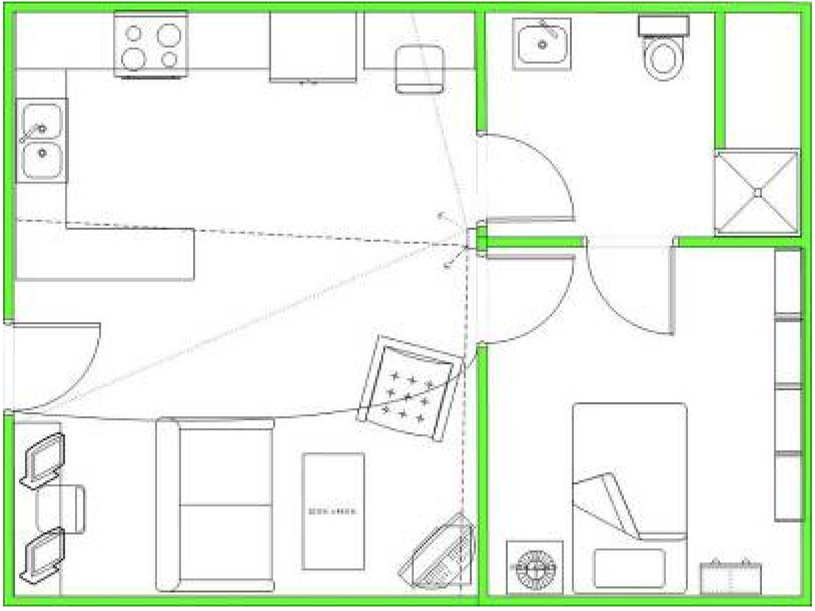

Figure 10: PoCL space with dotted lines to indicate kitchen camera field of view, a dotted arrow to indicate where the kitchen camera lens was pointed, dashed lines to indicate living room camera field of view, and a dashed arrow to indicate where the living room camera lens was pointed.

To achieve initial recordings indicating the conditions and to facilitate discussion about the equipment most suited for this situation, camera and stereo microphone pairs were pointed at each room. The distance between microphone centers on each stereo bar was approximately 15.2 cm on a *König & Meyer 23550 microphone bar* which was mounted to a boom microphone stand. The cameras were mounted independently on standoffs from the wall placed in the middle of the two microphones. The front and center of the living room camera lens stood off 18.1 cm from the wall. The front and center of the kitchen camera lens stood off 21.6 cm from the wall.

The microphones used were *CAD Professional CM217 cardioid condenser microphones* (stock photo, pickup pattern, and frequency response below in Figures **11** and **12**). They were chosen based on their pickup pattern and a frequency response within the vocal and percussive range. The center of the living room microphones was at a height of 182.2 cm. The center of the kitchen microphones was at a height of 185.4 cm. Both sets were angled downward toward the center of each room as depicted in Figure **10** by the arrows. Each microphone pair was powered by a *M-Audio Audio Buddy* dual microphone preamplifier (preamp). Finally, the outputs of the preamp were combined into one stereo 1/8" plug that fed into the line-in of an *Acer Aspire 3102WLMi* laptop where the sound was recorded using the open source audio recording software *Audacity*.

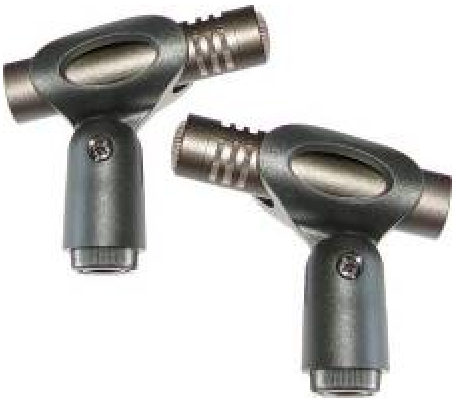

Figure 11: CAD CM217 Microphones.

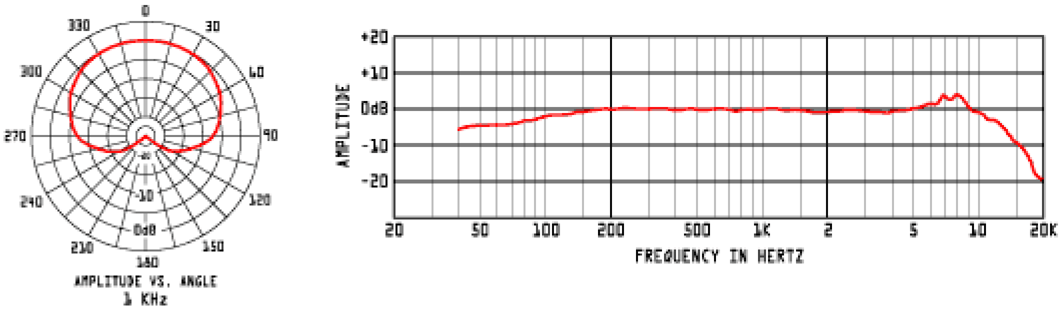

Figure 12: CAD CM217 Pickup Pattern and Frequency Response.

The cameras used were Unibrain Fire-i digital board firewire cameras (stock photo below in Figure **13**). The lenses chosen were relatively wide-view lenses with a focal length of 2.1 mm and a horizontal view angle of 80.95 degrees. Their video was collected using the provided *Unibrain Fire-i* software which records the video into an uncompressed .AVI file for each camera.

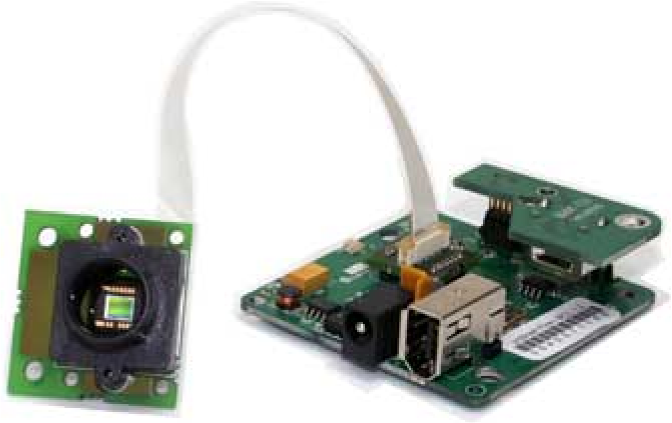

Figure 13: Unibrain Fire-i Digital Board Camera.

9. OHSU recording session I

The first recordings were created with the idea of capturing normal activities of daily living with the occasional fall or stumble as an abnormal event. The three scenarios chosen were:

- Walking from the couch to the sink, doing two dishes, and walking back.
- Walking from the couch to the fridge, looking for food, and walking back.
- Walking with a cane from the chair to the sink, washing hands, and walking back.

Figure 14 shows an individual frame of one of the recordings made with the OHSU original system.

These scenarios were captured collecting four normal takes, two stumble takes, and two fall takes per scenario for eighteen takes total.

The recordings were provided to the partners for discussion. Based on these recordings it was decided that the conditions allowed most of the AWEAR II hardware to be used in the PoCL. Using identical hardware as used in the AWEAR would allow the same recording software to be used, facilitating the exchange of data and preprocessing software.

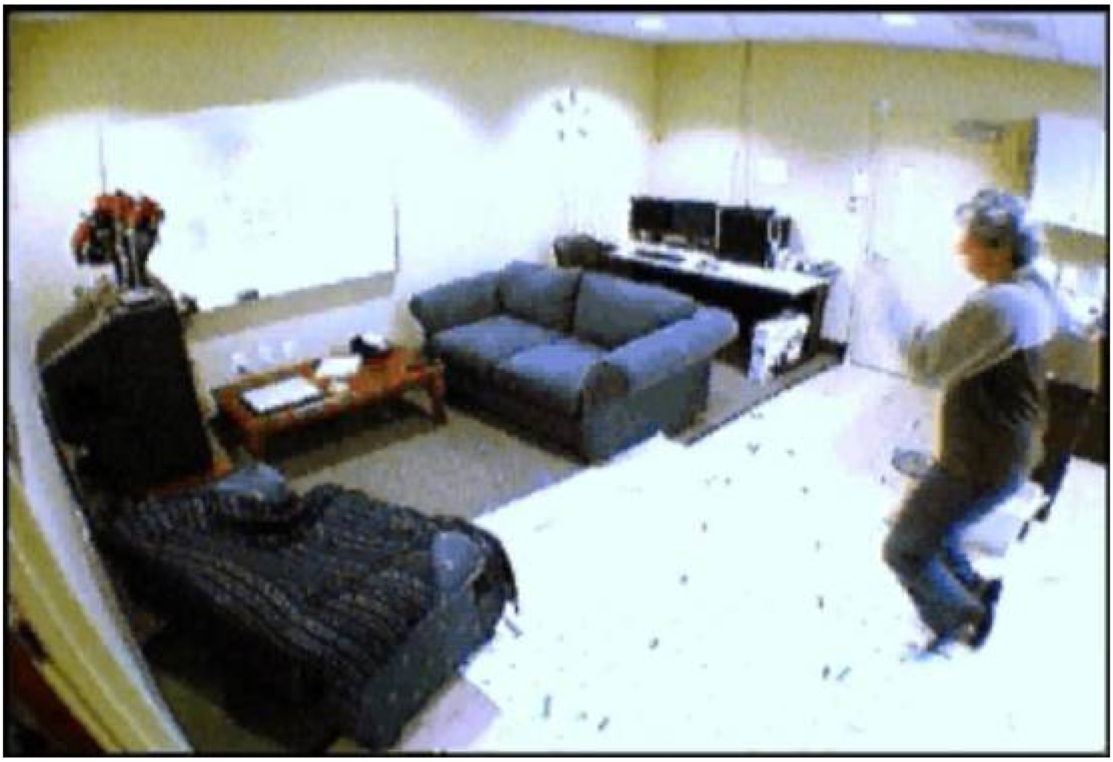

Figure 14: Image from initial recordings at OHSU, showing a person walking to the couch and stumbling.

10. Conclusion

For both the security and the in-home care scenario first recordings were made. These recordings were provided to the partners who have tested some of the methods developed on the images and sound data and provided valuable feedback on camera positioning and calibration, viewing angles, acoustic conditions and the implementation of the scenarios. This was especially important for the outdoor recordings with the moving AWEAR II platform, where conditions are less controllable and more variable.

The AWEAR II has been developed into an easy-to-use and very flexible recording platform with an autonomy of several hours.

11.Reference

Anemüller, J., Bach, J.-H., Caputo, B., Havlena, M. Jie, L., Kayser, H., Leibe, B., Motlicek, P., Pajdla, T., Pavel, M., Torii, A., Gool, L. v., Zweig, A. and Hermansky, H. "The DIRAC AWEAR Audio-Visual Platform for Detection of Unexpected and Incongruent Events", Proc. International Conference on Multimodal Interaction (ICMI) 2008, pp. 289-293.

Hengel, P. v. and Andringa, T. „Verbal Aggression detection in complex social environments" Proceedings of the 2007 IEEE International Conference on Advanced Video and Signal based Surveillance. 2007.

Hengel, P. v. and Anemüller, J. "Audio Event Detection for In-Home Care" Proceedings of the NAG-DAGA 2009 International Conference on Acoustics. 2009.