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.))

Project no: 027787

DIRAC
Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST – Priority 2

DELIVERABLE NO: D6.12
Catalogue of basic scenes containing incongruent events

Date of deliverable:30.06.2010
Actual submission date: 12.08.2010

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: FRA

Revision [0]

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.))

D6.12 - Catalogue of Basic Scenes Containing Incongruent Events

Fraunhofer Institute of Digital Media Technology, Project Group Hearing, Speech and Audio Technology (FRA)

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: 1. the security and surveillance domain, 2. the in-home monitoring of elderly people. For both application domains, scenarios have been developed to demonstrate the application of the methods developed in DIRAC. Various data has been recorded (D6.9, D6.10) to be used for testing and validation of the DIRAC methods. Furthermore, the recorded data was revised in D6.11 to investigate the necessity for additional data and to provide a strategy for testing and validation of the proposed DIRAC methods. The purpose of this catalogue is to provide a basic and atomistic testbed to the project partners in order to validate methods for rare and incongruent event detection. The recording equipment and setup is defined in detail to minimize the influence of errors that might affect the overall quality of the recordings in a negative way. Additional metadata, such as a defined format for scene descriptions, comments, labels and physical parameters of the recording setup is presented as a basis for evaluation of the utilized multimodal detectors, classifiers and combined methods for rare and incongruent event detection. The recordings presented in this work is available for the project partner via the DIRAC Wiki page [1] and will be made available online on the DIRAC project website [2].

Table of Content

1 Introduction

Within this project, scenarios have been developed and example situations have been recorded to show the potential of the DIRAC theoretical framework [3], while attempting to address realistic and interesting situations that can not be handled properly by existing technology. The methods show promising results on first recordings of near-real-life situations, but additional data is necessary to provide a strategy for testing and validation of those methods within the different scenarios of an application domain [4-5].

Two application domains are defined for the DIRAC project, namely the security market with its high demand for automated and intelligent surveillance systems, and the in-home care market with its need for monitoring elderly people at home. It was concluded that both domains would benefit considerably from the technology developed in the DIRAC project. In both domains there is a need for 24/7, unobtrusive, autonomous, and therefore intelligent, monitoring systems to assist human observers. Spotting and properly responding to unforeseen situations and events is one of the crucial aspects of monitoring systems in both application domains (see D6.1) [10].

The main purpose of this deliverable - as already mentioned in in D6.11 [6] – is to assemble a catalogue of recordings of very basic, atomistic scenes containing incongruent events for testing and validation of the DIRAC methods under a very controlled environment. Very basic in this context means for example only one person, only one action/incongruency, indoors, no cast shadows, enough light and defined sound sources. To support the idea of a controlled environment, the scenes and the necessary restrictions are defined in detail. Thus, it will be easier to compose more complex scenarios from atomistic scenes without having too much uncertainty about what can be processed with the detectors and classifiers forming the general and specific models within a DIRAC method.

The deliverable is organized as follows. First of all, the recording equipment and set is defined in detail in order to avoid errors and artifacts which would reasonably lead to unwanted results in individual classifiers and DIRAC methods. This includes a description of the AWEAR-II recording system in Section 2.1, information on the Communication Acoustic Simulator (CAS) recording room in Section 2.2 and the recording setup itself in Section 2.3. Afterwards, potential sources of errors are investigated and used to define characteristics of suitable audio-visual recordings in Section 3 and Section 4. The catalogue of basic audio-visual scenes is presented in

Section 5, followed by information on the format of the audio-visual recordings, their labels, scene descriptions and additional metadata in Section 5.2 and 5.3.

2 The recording environment

2.1 AWEAR-II recording platform

The AWEAR-II recording platform is a portable audio-visual recording platform that was developed in the DIRAC project (see D6.1, D6.6, D6.9, [9-11]). It consists of three mini-pcs (Siemens D2703-S mini ITX boards), two cameras (AVT Stingray), four microphones (T.Bone EM700 stereo microphone set), an audio capturing interface including a triggering device for audio-visual synchronization (FOCUSRITE Saffire PRO 10), a battery pack (Camden 12V gel) and a power distribution box. The hardware is mounted on a wearable backpack frame and allows human-centered recordings in both indoor and outdoor environments [7].

The AWEAR-II system can be controlled using a graphical user interface (GUI) application running on a netbook which in turn communicates with the recording platform via a wireless network connection. The GUI is used as a remote control for the preparation of recording sessions and capturing, for hardware adjustments and controlling. The recording data is stored on 2,5" removable hard disks connected to the AWEAR-II. After a recording session, the data is copied from the disks to a dedicated PC for further processing. For this purpose, format conventions have been defined. A more detailed description of the recording software as well as the processing steps has already been given in D6.9 [9] and will not be repeated here.

2.2 Communication and Acoustics Simulator – CAS

The Communication and Acoustics Simulator (CAS) at the House of Hearing in Oldenburg is a special room with variable acoustics that uses sophisticated techniques consisting of countless microphones, loudspeakers as well as large-scale electronics to create almost any acoustic room condition desired [8]. The CAS is usually used to run subjective tests on the latest hearing aids under various acoustic conditions or to test mobile phones for speech recognition and intelligibility in

realistic environments, but is also interesting for our work since environment dependent parameters can be manually controlled.

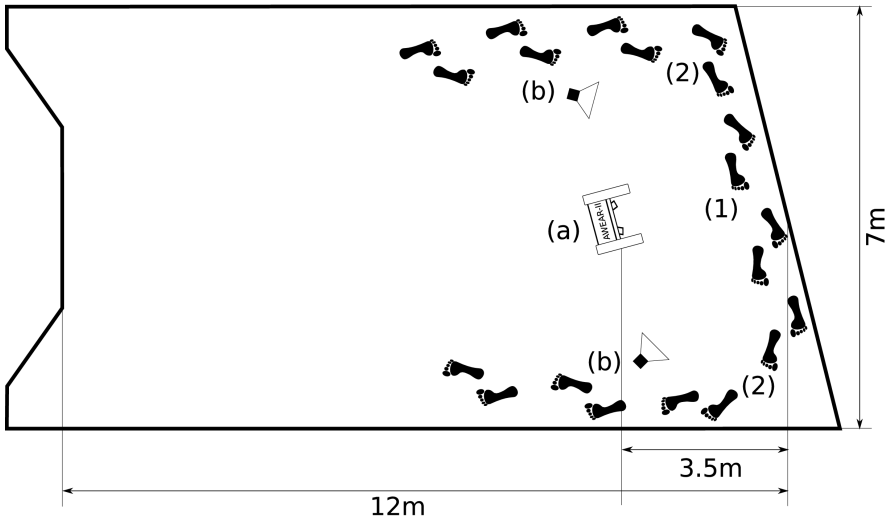

Figure 1: Floor plan of the CAS and the proposed recording setup. (a) denotes the location of the AWEAR-II recording system, (b) is the location of the additional light sources, (1) is defined as the center of the cameras field of view at a distance of approximately two meters and (2) are additional markers for the actors walking paths.

For parts of the recordings, the variable acoustics of the CAS was set up to the living room scenario with a reverberation time T60 of around 0.453 seconds. A floorplan of the CAS as well as additional information about the camera location and walking paths can be found in Figure 1. Additionally, a picture from the recording set is given in Figure 2.

Figure 2: A picture of the recording setup in the CAS.

2.3 Recording Setup

In this section, the scenes and the necessary restrictions are defined in more detail. The actors use a defined path to enter the scenery, to go trough the scenery and to leave it. Therefore, way points and walking paths have been defined and marked on the floor of the CAS in order to ensure repeatability of the experiments and and recordings. The AWEAR-II was placed orthogonal to the back wall in a distance of three and a half meters, whereas the mentioned way points and walking paths are in parallel to the wall in the back with a distance of approximately two meters from the camera center. This distance was found to be suitable enough to capture a wide viewing angle. All the parameters have been summarized and illustrated in the CAS floor plan as it can be found in Fig. 1.

All the recordings were made during daytime. However, to be independent from weather situations, i.e. constant illumination, two additional light sources were added to the scene. The consist of 400W spots attached to a stand and a dimmer to

control the amount of light introduced to the scene. Diffuse light situations are generated by using semi-transparent light shields which are attached to the spots.

The variable acoustics of the KAS was set up to the living room scenario. To illustrate the acoustic properties of the KAS at this setup, an estimation of the room impulse response is shown in Figure 3, where the reverberation time T60 was calculated to be 0.45298 seconds.

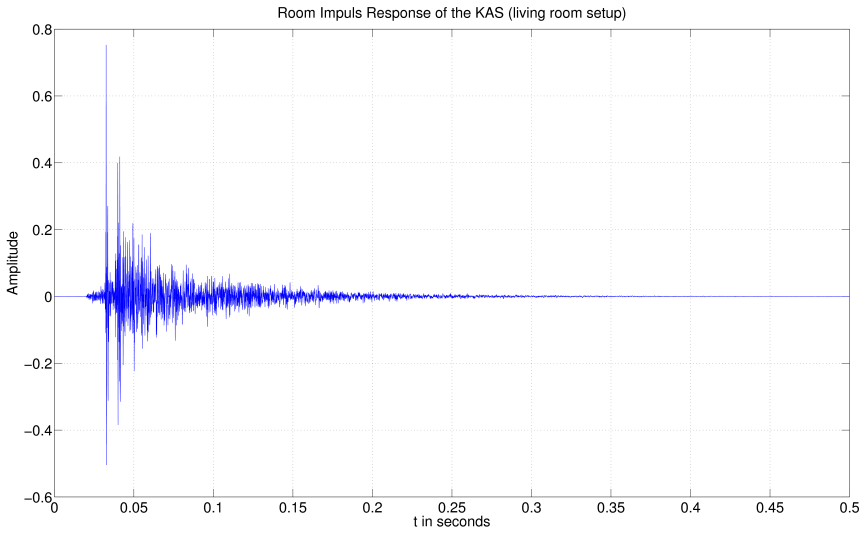

Figure 3: Room impulse response of the KAS (living room setup)

3 Reported problems with the audio-visual recordings

A large number of A/V sequences have already been recorded both for the in-home-care environment scenario and security surveillance (see D6.6 – D6.10). At the moment, 33 recording sessions containing more than 230 individual sequences are available on the DIRAC Consortium Server [2]. With such an amount of data to be evaluated, it is to be expected that not all of the detectors / models render the data

without failure.

Providing high quality audio-visual recordings for testing and evaluation is a non-trivial task. There is more to it than to define an audio and video format and physical parameters, such as the video frame rate and the audio sampling rate for the recording setup only. The recording environment needs to be defined as well in order to provide the best possible audio-visual quality. Therefore, it is crucial to identify potential problems and artifacts beforehand in order to completely remove or at least minimize their influence on the quality.

The main sources of errors and artifacts in general can be found in the recording environment and in a wrong calibration of the equipment. In particular, problems with improper illumination and cast shadows are unnecessarily common as well as problems with foreground-background separation as a consequence of improper dressing of the actors. The consequences are misclassification and confusion when running video based detectors and classifiers. Blurring as a result of defocussing, wrong gamma settings, shutter and white imbalance are problems in this context as well.

For audio recordings, the presence of unwanted sound sources, reverberations, humming, noise introduction and too low recording levels (low SNR) are serious problems, since they heavily influence the performance of audio based detectors and the overall quality of modality-fused detectors. Furthermore, improper synchronization between the audio and video data can be seen as a source of errors, too.

4 Evaluation of criteria for suitable audio-visual recordings

Motivated from the information on potential errors and artifacts reported by the project partners and as described in the previous section, the following requirements and criteria to create basic and atomistic audio-visual scenes containing incongruent events have been developed.

A suitable audio-visual sequence is defined as a sequence that allows error free processing of the DIRAC methods. Therefore, special attention needs to be drawn on sufficient illumination of the scene to minimize the influence of CCD camera chip

noise and motion blurring artifacts, especially when working with low video frame rates and fast moving objects. Cast shadows should be avoided to reduce false results/alarms from the video detectors. Furthermore, a high contrast between foreground objects and a steady, homogeneous background is desirable in order to avoid misclassification and confusion within the automated processing stages.

For the audio part, a suitable recording level should be selected such that the SNR stays reasonably high. Despite that the audio based detectors used in the DIRAC project have been shown to be robust against noise, the presence of uncontrollable noise sources should be minimized.

5 New basic audio-visual recordings for incongruency detection

5.1 Overview

Within the DIRAC project, keywords have been defined to describe the audio-visual recordings in the DIRAC in-home-care and security surveillance scenarios [Reference missing]. This was done for two reasons: First, to allow a search for audio-visual scenes within a database based on the keywords and second, to form complexity scalable scenarios by combining keywords. At the moment, six keyword groups exist. They are Movements, Interactions, Audio and their corresponding incongruent versions. The content of the keyword groups can be seen in Table 1. The covered keyword members are: standing, sitting, lying, walking, limping, stumbling, falling, backwards, one to N persons, speech, shouting, OOV and noise. In this recording session, only the indoor case is addressed.

Table 1: Keyword groups, its members and incongruencies available in the catalogue

Group	Members	Incongruencies
Movements	Standing, sitting, running, walking, lying, hesitating	Limping, stumbling, falling, backwards, sidewards, fleeing

Interactions	one to *N* persons, persons interact, dialog	fighting
Audio	Speech, noise	Monologue, shouting, out of vocabulary

In total, 100 audio-visual scenes have been recorded, containing samples to test individual detectors and classifiers for moving objects, visual foreground and background detection, person detection and localization, voice activity detection, speech recognition and the detection of acoustic events. Furthermore, basic audio-visual scenes have been recorded which cover incongruencies and rare events such as the ones given with the keywords. Thus, a verification and evaluation of individual detectors and classifiers is possible as well as verification and evaluation of the more complex DIRAC methods. Of course, the recordings are not only limited to the DIRAC methods, but can be used with any other approach.

5.2 List of new basic audio-visual recordings

Location: CAS (FRA, "Haus des Hörens", Oldenburg)

Setting: CAS Living Room Configuration

Recording platform: AWEAR 2.0 (fixed)

Framerate: 12 fps

Table 2: List of new basis audio-visual recordings

Scene	Actors	Equip.	Short Description
1	DH	-	walking from left to right in parallel to the camera axis, talking

2	DH	-	walking from right to left in parallel to the camera axis, talking
3	DH	-	walking backwards from left to right in parallel to the camera axis
4	DH	-	walking backwards from right to left in parallel to the camera axis
5	DH	-	walking sidewards, frontal view, from left to right in paralel to the camera axis
6	DH	-	walking sidewards, frontal view, from right to left in paralel to the camera axis
7	DH	-	walking sidewards, back view, from left to right in paralel to the camera axis
8	DH	-	walking sidewards, back view, from right to left in paralel to the camera axis
9	DH	-	walking from left to right to the center, turning around (facing camera) and leaving
10	DH	-	walking from left to right to the center, turning around (facing wall) and leaving
11	DH	-	walking from right to left to the center, turning around (facing camera) and leaving
12	DH	-	walking from right to left to the center, turning around (facing wall) and leaving
13	DH	chair, sofa, table, ...	walking from left to right in parallel to the camera axis, crossing an object in the foreground that covers parts of the body
14	DH	chair, sofa,	walking from right to left in parallel to the camera axis, crossing an object in the foreground

		table, ...	that covers parts of the body
15	DH	chair, sofa, table, ...	walking from left to right in parallel to the camera axis, crossing an object in the foreground and hiding completely
16	DH	chair, sofa, table, ...	walking from right to left in parallel to the camera axis, crossing an object in the foreground and hiding completely
17	DH	-	limping from left to right in parallel to the camera axis
18	DH	-	limping from right to left in parallel to the camera axis
19	DH	-	limping from left to right to the center, continue with walking normal
20	DH	-	limping from right to left to the center, continue with walking normal
21	DH	-	walking from left to right to the center, continue with limping
22	DH	-	walking from right to left to the center, continue with limping
23	DH	something to pick up	walking from left to right to the center, picking something up, continue with normal walking
24	DH	something to pick up	walking from right to left to the center, picking something up, continue with normal walking
25	DH	something to pick up	walking from left to right to the center, picking something up, turning around and leaving

26	DH	something to pick up	walking from right to left to the center, picking something up, turning around and leaving
27	SG	-	walking from left to right to the center, stumbling, continue with normal walking
28	SG	-	walking from right to left to the center, stumbling, continue with normal walking
29	SG	-	walking from left to right to the center, stumbling, turning around and leaving
30	SG	-	walking from right to left to the center, stumbling, turning around and leaving
31	SG	-	walking from left to rightto the center, stumbling, moaning, continue with limping
32	SG	-	walking from right to left to the center, stumbling, moaning, continue with limping
33	SK	-	walking from left to right to the center, falling, moaning, standing up, continue with normal walking
34	SK	-	walking from right to left to the center, falling, moaning, standing up, continue with normal walking
35	SK	-	walking from left to right to the center, falling, moaning, standing up, turning around and leaving (take a: stand up first, take b: standing up and turning at once)
36	SK	-	walking from right to left to the center, falling, moaning, standing up, turning around and leaving (take a: stand up first, take b: standing up and turning at once)

37	SK	-	walking from left to right to the center, falling, moaning, standing up, continue with limping
38	SK	-	walking from right to left to the center, falling, moaning, standing up, continue with limping
39	DH	chair, sofa, table, ...	walking from left to right to the center, sitting down (chair, sofa, floor), frontal view, standing up and continue walking
40	DH	chair, sofa, table, ...	walking from right to left to the center, sitting down (chair, sofa, floor), frontal view, standing up and continue walking
41	DH	chair, sofa, table, ...	walking from left to right to the center, sitting down (chair, sofa, floor), back view, standing up and continue walking
42	DH	chair, sofa, table, ...	walking from right to left to the center, sitting down (chair, sofa, floor), back view, standing up and continue walking
43	SG	chair, sofa, table, ...	walking from left to right to the center, sitting down (chair, sofa, floor), frontal view, falling down, moaning, standing up and continue walking
44	SG	chair, sofa, table, ...	walking from right to left to the center, sitting down (chair, sofa, floor), frontal view, falling down, moaning, standing up and continue walking
45	SG	chair, sofa, table, ...	walking from left to right to the center, sitting down (chair, sofa, floor), back view, falling down, moaning, standing up and continue walking
46	SG	chair, sofa,	walking from right to left to the center, sitting down (chair, sofa, floor), back view, falling

		table, ...	down, moaning, standing up and continue walking
47	SG	chair, sofa, table, ...	walking from left to right to the center, sitting down (chair, sofa, floor), frontal view, falling down, moaning, standing up and limping
48	SG	chair, sofa, table, ...	walking from right to left to the center, sitting down (chair, sofa, floor), frontal view, falling down, moaning, standing up and limping
49	SK	chair, sofa, table, ...	walking from left to right to the center, sitting down (chair, sofa, floor), back view, falling down, moaning, standing up and limping
50	SK	chair, sofa, table, ...	walking from right to left to the center, sitting down (chair, sofa, floor), back view, falling down, moaning, standing up and limping

5.3 Format of the Audio-visual Recordings

In order to provide the best possible quality to the research community, all the recordings are stored in an uncompressed data format. In particular, the Portable Network Graphics (PNG) format is used camera channel-wise on a frame by frame basis with a 1920x1080 resolution. For the four microfon channels, all recordings are stored as wav-files with 48kHz sampling rate and a sample resolution of 32 bit floating-point.

5.4 Metadata and Scence Descriptions

Each audio-visual recording contains a label file which includes information about the name of the recording and the scene, the date, the location, the used device, frame rate, a placeholder for comments, as well as a detailed description of the scene. All the labels have been generated manually, either by hand or by using custom made semi-supervised tools. Optionally, the audio-visual scene is rendered as a preview video using one or both camera signals and the front stereo microfon set.

For this purpose, any video codec can be used since this is done only for preview purposes. In combination with the additional metadata and scene descriptions as exemplary given in Table 3, the selection of suitable audio-visual recordings is facilitated. The metadata has been stored together with the audio-visual recordings on the DIRAC project webpages [1-2].

Table 3: Example of a label file to provide additional and contextual information about an audio-visual recording.

```
# Recording: AggressionRecordings
# Scene : AggressionScenes
# Date : 20091216
# Location : house front next to HdH (FRA Oldenburg)
# Equipment: AWEAR-II (fixed)
# Framerate: 12 fps
# Comments : shades visible

# start time (in sec) | end time (in sec) | key words | short
  description;
0000 | 0001 | persons 2, walking | Two persons walk towards each
               other;
0002 | 0003 | persons 2, walking, speech | Defensive person begins
               conversation;
0003 | 0005 | persons 2, persons interact, shouting | Aggressor
               suddenly starts fighting;
0005 | 0010 | persons 2, falling, fleeing, shouting | Person breaks
               down, aggressor flees;
0010 | 0017 | persons 1, limping | Person limps away;
...
```

6 Conclusion

A catalogue of basic and atomistic audio-visual recordings for rare and incongruent event detection was presented in this paper. While adressing the work within the ongoing DIRAC project at first, the applicability of the catalogue is not limited to the methods developed there. A careful definition and analysis of the recording environment, the recording equipment and setup is seen to be crucial for two reasons. First, to provide the best audio-visual quality of the recordings

achievable to ensure that the performance of utilized detection schemes and classifiers do not degrade with quality of the test data. Second, to focus on the validation and evaluation of novel combined detection schemes and modality-fused event detection methods such as the ones proposed in DIRAC instead of the underlying algorithms for modeling only single information instances. The catalogue will be used to validate the methods for rare and incongruent event detection developed within the DIRAC project and will probably be extended in the future based on the needs of the project partners. Both the recordings and the results are available on the project websites [1-2].

References

[1] DIRAC Project Wiki Page: https://dirac.uni-oldenburg.de/DIRAC

[2] IST-027787 project website: Detection and Identification of Rare Audio-visual Cues - DIRAC. http://www.diracproject.org/

[3] Weinshall D., Hermansky H., Zweig A., Luo J., Jimison H., Ohl F.,and Pavel M.: *Beyond Novelty Detection: Incongruent Events, when General and Specific Classifiers Disagree*. Advances in Neural Information Processing Systems (NIPS), Vancouver, December 2008, (2008)

[4] Hengel, P.W.J. van, Andringa, T.C.: *Verbal aggression detection in complex social environments*. In Proceedings of AVSS 2007, (2007)

[5] Hengel, P.W.J. van, and Anemüller, J.: *Audio Event Detection for In-Home-Care*. In NAG/DAGA International Conference on Acoustics, Rotterdam, 2326 March 2009, (2009)

[6] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.11 "Testing and evaluation plan"

[7] Havlena, M., Ess, A., Moreau, W., Torii, A., Janoek, M., Pajdla, T., Van Gool, L.: *AWEAR 2.0 system: Omni-directional audio-visual data acquisition and processing*. In: EGOVIS 2009: First Workshop on Egocentric Vision. pp. 4956 (2009)

[8] Behrens, T.: *Der 'Kommunikationsakustik-Simulator' im Oldenburger Haus des Hörens*. 31. Deutsche Jahrestagung für Akustik: Fortschritte der Akustik DAGA 2005 (1), München, DEGA e.V., pp. 443–445, (2005)

[9] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.9 "Database of Recordings of Scenario 1"

[10] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.1 "Application Scenarios"

[11] DIRAC – Detection and Identification of Rare Audio-visual Cues, Deliverable D6.6 "First audio-visual data collection with AWEAR and OHSU System"