# Catalog of Basic Scenes for Rare/Incongruent Event Detection

Danilo Hollosi[1], Stefan Wabnik[1], Stephan Gerlach[2], and Steffen Kortlang[2]

[1] Fraunhofer Institute for Digital Media Technology,
Project Group Hearing Speech and Audio Technology, D-26129 Oldenburg, Germany
Danilo.Hollosi@idmt.fraunhofer.de,
WWW home page: http://www.idmt.fraunhofer.de/eng
[2] Carl von Ossietzky University of Oldenburg
Institute of Physics
D-26111 Oldenburg
Germany

**Abstract.** A catalog of basic audio-visual recordings containing rare and incongruent events for security and in-home-care scenarios for European research project *Detection and Identification of Rare Audio-visual Cues* is presented in this paper. The purpose of this catalog is to provide a basic and atomistic testbed to the scientific community in order to validate methods for rare and incongruent event detection. The recording equipment and setup is defined to minimize the influence of error 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 are available online on the DIRAC preoject website [1].

**Keywords:** Rare event detection, Incongruency, Databases

## 1 Introduction

The performance of state-of-the-art event detection schemes usually depend on a huge amount of training data to be able to identify an event correctly. The less data is available, the more problematic the abstraction of the event to a model becomes. This is especially true for rare event detection, where events have a very low a-priori probability. In the same context, this is also true for incongruent events. They are defined as rare events which conflict with commonly accepted world models. Consequently, novel methods need to be developed to detect such events and to compensate for the lack of suitable training data.

Possible solutions for this problem are currently investigated in the European research project *Detection and Identification Of Rare Audio-visual Cues* (DIRAC). The main idea of the DIRAC approach for event detection is - in

contrast to existing holistic models - to make use of the discrepancy between more general and more specific information available about reality [2]. This approach heavily reduces the amount of training data necessary to model rare and incongruent events and can be generalized in such a way that information from different modalities become applicable as well.

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.

For both of them, scenarios have been developed and example situations have been recorded to show the potential of the DIRAC theoretical framework, 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.

This paper presents a catalog of recordings of very basic, atomistic scenes containing incongruent events that are suitable for testing and validation of the DIRAC methods. 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 underlying, utilized detectors and classifiers forming the general and specific models within a DIRAC method.

This work 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 2.4. To the heart of this work, the catalog of basic audio-visual scenes is presented in section 3, followed by information on the format of the audio-visual recordings, their labels, scene descriptions and additional metadata in section 3.1 and 3.2.

## 2 Methods

### 2.1 AWEAR-II Recording System

The AWEAR-II recording platform is a portable audio-visual recording platform that was developed in the DIRAC project. It consists of three mini-pcs (*Siemens D2703-S mini ITX boards*), two cameras (*AVT Stingray*), four microphones (*Tbone EM700 stereo mic 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 [4]. A picture of the AWEAR-II recoding system can be found in Fig. 1.

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 will be described in section 3.1 and section 3.2.

**Fig. 1.** The AWEAR-II recording system with nettop remote control

### 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 [6]. 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.

For part 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 floor-plan of the CAS can be found in Fig. 2, whereas additional information about the camera location and walking paths can be found there as well.

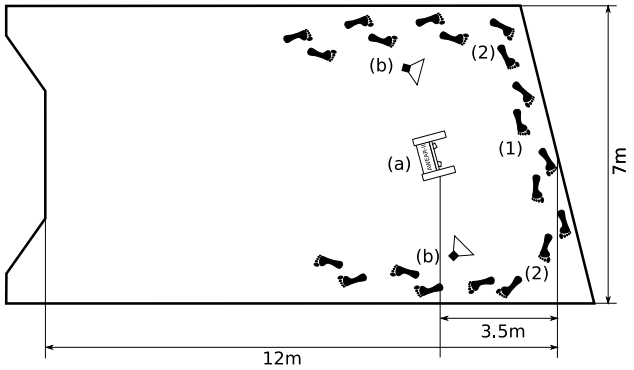


**Fig. 2.** Floor plan of the CAS and the proposed recording setup. *(a)* denotes the location of the AWEAR-II recording system, *(b)* is the location of the additional light sources, *(1)* is defined as the center of the cameras field of view at a distance of approximately two meters and *(2)* are additional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

### 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. 2.

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 100W 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.

### 2.4 Potential Sources of Errors

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 problem since they heavily influences 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.

### 2.5 Evaluation criteria for audio-visual recordings

Motivated from the information on potential errors and artifacts 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.

## 3 Basic Scenes for Incongruency Detection

Within the DIRAC project, keywords have been defined to describe the audio-visual recordings in the DIRAC in-home-care and security surveillance scenarios. 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, 6 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.

**Table 1.** Keyword groups, its members and incongruencies available in the catalog

| **Group** | **Members** | **Incongruency** |
|---|---|---|
| Movements | standing | limping |
| | sitting | stumbling |
| | running | falling |
| | walking | backwards |
| | lying | sidewards |
| | hesitating | fleeing |
| Interactions | one to N persons | fighting |
| | person interact | |
| | dialog | |
| Audio | speech | monolog |
| | noise | shouting |
| | | out-of-vocabulary |

In total, 95 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 recording are not only limited to the DIRAC methods, but can be used with any other fused modality approach.

### 3.1 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 float.

### 3.2 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 given in Fig. 4, selection of suitable audio-visual recordings is facilitated. The Metadata has been stored together with the audio-visual recordings on the DIRAC project page www.diracproject.org [1].

```
# 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;
...
```

**Fig. 3.** Example of a label file to provide additional and contextual information about the audio-visual recording

## 4 Conclusion

A catalog 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 catalog 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 catalog 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 will be published on the project website [1].

## 5 Acknowlegdement

The authors would like to thank *Haus des Hörens* for providing access to the CAS, technical support and equipment. This work was supported by the European Commission under the integrated project DIRAC (Detection and Identification of Rare Audio-visual Cues, IST-027787).

## References

1. IST-027787 project website: Detection and Identification of Rare Audio-visual Cues - DIRAC. http://www.diracproject.org/
2. 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)
3. Hengel, P.W.J. van, Andringa, T.C.: Verbal aggression detection in complex social environments. Proceedings of AVSS 2007, (2007)
4. 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)
5. Hengel, P.W.J. van, and Anemüller, J.: Audio Event Detection for In-Home-Care. NAG/DAGA International Conference on Acoustics, Rotterdam, 2326 March 2009, (2009)
6. 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., 443–445 (2005)