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: D6.11
Testing and validation plan

Date of deliverable: 30.05.2010
Actual submission date: 29.06.2010

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.))

D6.11 – TESTING AND VALIDATION PLAN

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 to be used for testing and validation of the DIRAC methods.

The purpose of this deliverable is to list the data already collected by the partners, as well as additional scenes still needed, and the experiments developed by the partners to which the data has been or will be applied to. Based on the plans of each partner and the tasks still to be done, a detailed testing and evaluation plan will be given.

Table of Content

1 Introduction

In the Technical Annex submitted at 12th of February 2009, two application scenarios were defined for the DIRAC technology, namely the security market with its high demand for automated and intelligent surveillance systems (scenario 1), and the in-home care market with its need for monitoring elderly people at home (scenario 2). 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. 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 to show the potential of the DIRAC theoretical framework and the techniques developed in the various work packages, while attempting to address realistic and interesting situations that can not be handled properly by existing technology. To make these developed methods and techniques in DIRAC applicable in both scenarios, and to learn about their capabilities and restrictions, example situations have been recorded for both scenarios. The recordings have been sorted into databases (D6.6, D6.7, D6.8, D6.9, and D6.10).

Apart from the audio-visual recordings, the partners collected data for the development of the detectors. Different combinations of detectors have been used in several experiments.

In what follows, the experiments already done by the partners as well as planned experiments are listed together the relevant data bases used for the experiments (see sections 2 an 3). Thereafter, validation methods as well as a validation plan are given (see section 4 and 5).

2 Experiments

The next two subsections present the experiments and their building blocks developed by the partners.

2.1 Detectors

In the following, a short description of the different Detectors already developed by the partners will be given. The detectors are the building blocks of the experiments.

2.1.1 Sound source localization

The sound source localization method is targeted to estimate the direction of arrival of sound sources which are active in an acoustic scene. The approach involves an analysis of correlation between different channels of the audio data with a subsequent classification step. The features used for classification as well as the final localization results of the audio-only analysis are provided for further application, e.g. in the fused AV-detector (see Deliverable 1.8).

Specifically designed recordings were obtained to calibrate the audio and video localization. This resulted in a unique conversion between spatial units native to audio and those native to video, which proved of value in all AV-related applications.

2.1.2 Acoustic object detection

Partner OL has developed algorithms to identify acoustic objects in various backgrounds. The method is based on analyzing the amplitude modulations in the sound signal and uses supervised learning to discriminate objects from background. In particular, speech/non-speech detection has been implemented, and detectors for several office objects are available (see Deliverables 2.5 and 2.11). The results of the classification are provided to all partners as binary decisions and confidences, to be used e.g. as additional information in the AV-analysis.

Many of the scenes recorded at OHSU and at FRA helped to adapt the algorithmic implementation. In particular, a high false hit rate on speech detection in some recordings led to a modified training stage, which improved the performance of the algorithm on recorded data.

2.1.3 *tracker tree*

ETHZ has developed tracker trees for the detection of known and unknown human actions (c.f. deliverable 3.9). The basic idea is to detect abnormal human motions from the hierarchical incongruence using a set of trackers that have different information about what they track in the 'normal world'. If a tracker at a certain location in the tracker tree fails, but other trackers still work well, abnormalities are detected and semantic interpretation on the nature of this abnormality is deduced. Additionally, inspired by the tracker trees, ETHZ proposed a technique to automatically learn a model of normal human behaviour where appearances and actions are described in two distinct hierarchies. In each hierarchy, abnormal event detection and reasoning is possible. This work has recently been published (Nater et al., CVPR 2010).

2.1.4 *Out-of-vocabulary word detection*

The out-of-vocabulary (OOV) word detection system developed by BUT is based on a hybrid word/sub-word recognizer. A specific part in the hybrid model of the recognizer covers the most frequent English words, whereas OOV words are covered as sequences of sub-words in the generic part of the model. The recognizer determines the most likely sequence of words and sub-words, hypothesizing potential OOV words and describing them as sub-word sequences in the recognition output. A detailed description of the system is given in Deliverable D2.12. Scenes recorded by OHSU were used to put the system to a hardness test as shown in the demonstrational video scene shown at this years review meeting. The recognition output, detected OOV words and confidence scores are to be provided at the final WP6 deliverable.

2.1.5 *Person detector*

The INRIA OLT detector toolkit based on the histograms of oriented gradients (HOG) algorithm presented by Dalal and Triggs is used to detect upright standing people in video input data. It scans an image using a sliding detection window technique to detect the presence or absence of a human-like shape at every possible position in the image. Obtained detection scores (confidences) are further processed using non-maxima suppression with robust mean shift mode detection in order to avoid multiple detections of the same person. Results of the visual-only analysis are provided for further application, e.g. in the fused AV-detector (see Deliverable 1.8).

2.1.6 General visual blob detectors

Several state-of the-art visual blob detectors, e.g. MSER, SIFT, and SURF, are used as feature image point detectors forming the first step of a structure from motion pipeline developed at CTU that allows for camera tracking and image stabilization w.r.t. the ground plane as needed by the person detector.

Statistics of the camera tracking can be further used to detect abnormal situations, i.e. "Camera fails", "Camera out of focus", "Cameras out of sync", and "Camera rig calibration wrong" (see Deliverable 3.8).

2.2 Experiments

This subsection lists the experiments already accomplished by the partners as well as planned experiments for Y5 of the project.

2.2.1 LIN

A main output of work in WP5 has been the identification of the principle of cross modal category transfer as a means to cope with a rare, incongruous event in one sensory modality. This principle, while being further investigated within WP5, has already been taken up by WP4 investigating possible implementations in engineering systems.

2.2.2 IDIAP, ETHZ

IDIAP plans to test its algorithms on knowledge transfer across concepts on the data acquired at ETHZ, imaging different subjects performing various types of actions. The goal will be to detect an n incongruent action and then learn it quickly exploiting previously learned models. This work will be carried out in collaboration with partner ETHZ.

2.2.3 BUT

So far, BUT has developed two systems for OOV word detection, which have been examined already earlier and are described in the deliverables D6.9 and D6.12. Now, partner BUT plans to concentrate their research on out-of-language detection. Necessary data will be mainly recorded in-house in a meeting scenario. Existing systems such as their language identification system and both of their OOV detectors could serve as base for the new system.

Out-of-Language detection can be regarded as a less specialized task of detecting unexpected acoustic input than OOV word detection.

2.2.4 CTU, OL, FRA

CTU developed a fused AV-detector which decides presence/absence of a speaker in the scene directly from audio and visual features (audio features provided by OL). Incongruent observations are detected when the decisions of the fused AV-detector differ from the conjunctions of the decisions of separate audio and visual speaker detectors. Several sequences were recorded by FRA in order to both train and validate the detector.

2.2.5 OL

OL has worked on acoustic scene analysis using localization and detection of sound sources. Our approach allows identifying acoustic objects and backgrounds, which has already been implemented in an audio-only rare event detection framework.
Combination of sound source localization and acoustic object identification will provide information about where and when a potentially rare event occurred in a known scenario.

2.2.6 KUL, ETHZ

The locomotion stimuli have been used in one published behavioural study (Vangeneugden, J., Vancleef, K., Jaeggli, T., Van Gool, L., Vogels, R. (2010). Discrimination of locomotion direction in impoverished displays of walkers by macaque monkeys. *Journal of Vision, 10* (4), 22.1-22.19) and in a set of single cell recording studies in monkey temporal cortex (see Deliverable 3.10. to be submitted June 2010).

2.2.7 FRA

FRA has developed a „conversation detector“ which - via incongruency detection - signals unusual conversation situations, e.g. a person talking to himself.

The conversation detector is composed using the parts-whole relationship. Data from the audio-visual domain is processed by three detectors, the person detector from the tracker tree (ETHZ), the sound source localisation (OL) and the speech/non-speech classificator (OL). Data fusion from all three detectors is processed by the specific model (SVM) to detect incongruency.

2.2.8 ETHZ

ETHZ will continue the use of tracker trees in the upcoming test scenarios (sequences recorded at FRA and OHSU). The recently developed technique with self-learning hierarchies will also be tested on all DIRAC scenarios, encouraging the more long term analysis of human behaviour in indoor surveillance scenes, as well as the aspect of updating the existing model. A further plan is to incorporate scene-specific knowledge in the detection of visual incongruencies.

3 Data Bases

In this section, an overview over the different data bases already assembled by the partners is given.

3.1 Data base of test recordings for AWEAR-beta

The data forming this data base has been recorded with several AWEAR prototypes and the AWEAR-beta device (see Deliverable 1.4). The data has been used for the evaluation of separate detectors (visual person detector, sound source localization, speech/non-speech classifier) and for the first AV-fusion experiment. Gender data has been used for IDIAP experiments.

The following tables list information about each recording using these abbreviations:

- date: date of recording,
- name: name of the recording session (as listed on the DIRAC wiki page)
- #: number of takes
- short description of the scene

date	**name**	#	**short description**
2007-10-17	Language Outdoor	18	Dialogue with AWEAR user in different languages, different distances, different angles (rare language)
2007-10-17	Talking Outdoor	11	Speaking while approaching AWEAR-beta user or passing by
2007-10-17	FromSide Outdoor	02	Person speaking from side (not visible) while a silent person is visible
2007-10-18	Language Indoor	16	Dialogue with AWEAR user in different languages, different distances, different angles

2007-10-18	Talking Indoor	14	Speaking while approaching AWEAR-beta user or passing by
2007-10-18	FromSide Indoor	08	Person speaking from side (not visible) while a silent person is visible
2007-12-07	Gender Congruent	17	Approaching AWEAR and speaking normally
2007-12-07	Gender Incongruent	11	Approaching AWEAR and speaking with a low/high pitch voice
2007-12-07	Gender Distorted	02	Approaching AWEAR and speaking with a distorted voice
2008-03-17	cell phone sequences	08	A person approaches AWEAR-beta walking normally or in a strange way searching for his lost cell phone. The person starts a conversation with the AWEAR-beta user. Then a third person approaches from a hardly visible angle and joins the conversation
2008-03-17	sequences with dialogues in different languages	15	A person approaches AWEAR-beta and asks the AWEAR-beta user for help in a native language. The AWEAR-beta user replies that he does not understand and the person repeats the same question in English, gets a reply, and the conversation continues in English.
2008-03-17	fire sequences	06	A person approaches AWEAR-beta and asks the AWEAR-beta user for help in a native language. The AWEAR-beta user replies that he does not understand and the person repeats the same question in English, gets a reply, and the conversation continues in English. The conversation is interrupted by a person storming into the room shouting "FIRE”.
2008-03-17	sequences with a moving camera	04	Same as before, but AWEAR-beta is moving

3.2 Data base of audio-visual recordings

The data base contains recorded scenes for the two scenarios 'home-care of elderly people' and 'surveillance'. The recorded scenes exemplify incongruencies in the audio, visual and audio-visual domain. One or more actors have been recorded outdoor and indoor with different activities (moving, walking, stopping, falling, talking, etc.). The recordings have been made utilizing the AWEAR-II recording platform, the OHSU setup and the ETHZ test setup.

The following tables list information about each recording using these abbreviations:

- date: date of recording,
- name: name of the recording session (as listed on the Dirac Wiki page)
- #: number of takes
- short description of the scene
- P: processed or not ('√' for processed, empty if not),
- Inc.: modality of the incongruency (A: audio, V: video, A\V incongruency between audio and video)

3.2.1 Scenario "surveillance", outdoor recordings

date	name	#	short description	P	Inc.
2008-11-24	BicycleFromBehind	05	Bicycle overtakes pedestrian ringing and shouting		V
2008-11-24	PedCrossStreet	04	Almost crash between pedestrian and bicycle	√	V
2009-02-05	WalkingBruno	20	Five different actors walking (1) , hesitating (2), falling (3) and running (4)	√	V
2009-03-18	PedestrianFalling	02	A Pedestrian falls with two other pedestrians passing by		V
2009-03-18	PedestrianFalling	02	A Pedestrian falls with two other pedestrians passing by		V
2009-07-14	CollisionCourse	32	Three different versions of an almost crash with a car/bicycle (video over-exposed)		V
2009-07-28	CollisionCourseRem	16	Remake/modification of the scene CollisionCourse	√	V

2009-08-26	Interview	01	Recording of improvised interview with oov words, busy city		V,OOV
2009-08-26	MovingObjects	01	People (walking, falling, passing), bicycles and cars passing		V
2009-09-10	OldenburgCity	06	Recordings of the inner city of Oldenburg		V
2009-12-16	StealingBook	06	Three different versions of a book-robbery		V
2009-12-16	AggressionScenes	07	Two persons hustle each other, roaring (different variations) variations)	√	V

3.2.2 Scenario "surveillance", indoor recordings

date	**name**	**#**	**short description**	**P**	**Inc.**
2009-06-11	WalkingAnteroom	08	Guys walking towards on each other, passing by or greeting		V
2009-12-08	WalkingStyles	15	Two guys walking in different variations		V
2009-12-08	SpeakerIdentification	08	Guys walking towards on each other, passing by or greeting		V
2009-12-08	StealingBag	02	Three different versions of a bag-robbery		V

3.2.3 Scenario "in-home care of elderly people", indoor recordings

date	**name**	**#**	**short description**	**P**	**Inc.**
2008-11-25	People walking	02	One person walking and talking all the time with loudspeaker speaking	√	A\V
2009-04-06	Review scene	01	Guy is walking and tem-porarily talking in presence of a loudspeaker.	√	A\V

2010-01-14	OOV Speech 2	02	Person walks around and talks on the phone (head set)		OOV
2010-01-19	Person falling 1	02	Person falling, calling for help (oov name)		V,OOV
2010-01-20	Person falling 2	02	Person falling, calling for help		V
2010-01-21	Person walking 1	02	Person walks carefully as if on ice		V
2010-01-22	Person walking 2	02	Person walks like an elder		V
2010-03-05	Person falling 3	04	Person enters the room, walks around, falls down, stands up, walks, leaves the room		V
2010-03-05	Person picking up something	02	Person enters, walks around, picks something up from the ground, walks around, leaves		V
2010-03-05	Person laying down	02	person enters, walks to the couch, lies on the couch, waits, stands up again, walks, leaves		V
2010-03-06	OOV Speech 2		person enters, walks around, starts talking, uses some oov words while walking, stops talking, walks, leaves		V, OOV
2010-03-17	Woman telephone	04	Person talking on the phone		OOV
2010-03-17	Knocking	03	Person in room asks knocking subject in		
2010-03-17	Enter room	05	Person entering and leaving the scenery through door and have a conversation with second person	√	A,V
2010-93-17	Radio active	04	Persons listening to an interview on the radio, uses radio with its panel	√	A\V
2010-03-17	Radio remote	03	Persons uses radio with remote control and listening to an interview on the radio	√	A\V
2010-03-17	Stand up and talks to itself	05	Person walking around while talking to itself	√	A\V
2010-03-17	Hit and limp	06	Persons hits couch and limps out of the image		V
2010-03-17	Walk behind couch	04	Persons walks through scenery without hurting itself		V

2010-03-17	Pick up	05	Person picks up something already on the floor		V
2010-03-17	Drop and pick up	04	Person drops a key and picks it up again		V
2010-03-17	Tie laces	04	Person ties his shoe laces		V
2010-03-17	Stumbling	04	Person stumbles upon the couch		V
2010-03-17	Falling down	04	Person stumbles and falls down completely		V
2010-03-17	Lie on couch	03	Person lies on the couch and calls for help		A,V
2010-03-17	Pick up and falling down	04	Person tries to pick something up, but breaks down		V
2010-03-17	Drop and falling down	03	Person drops a key, tries to pick it up again, but breaks down		V

Sound classification:

date	**name**	**#**	**Short description**	**P**	**Inc.**
2010-04-01	Office bird	04	Office scene with bird sound		A
2010-04-01	Office bird noise	01	Office scene with bird sound and additional noise		A
2010-04-01	Office telephone	05	Office scene with phone call		A
2010-04-01	Office telephone noise	01	Office scene with phone call and additional noise		A

3.2.4 Low-level incongruencies

date	**name**	**#**	**short description**	**P**	**Inc.**
2009-03-18	Parking lot with cam cover	01	Walking along a parking lot, when suddenly one camera fails	√	V
2009-04-06	PedestriansCamCover	01	One camera is temporarily covered while meeting pedestrians	√	V

3.2.5 *Low-complexity, indoor*

date	name	#	Short description	P	Inc.
2010-02-11	KASWalkChair	06	Person walking through image or sitting on chair		V

3.2.6 *Calibration scenes, indoor*

date	name	#	short description	P	Inc.
2008-11-25	People talking	04	Person speaks from defined angle	√	-
2008-11-25	Loudspeaker talking	04	Loudspeaker speaks from defined angle	√	-
2008-11-25	People shouting	04	Person shouts from defined angle	√	-
2009-01-19	Loudspeaker Calibration	13	Calibration scenes with a loudspeaker (13 positions)	√	-
2009-01-19	Walking the line	05	Person walking three times left to right and back along a line	√	
2009-01-19	Walking with noise	04	… with an additional loudspeaker at a fixed position	√	
2009-01-19	Person vs. Loudspeaker	05	People and loudspeaker at fixed positions cross-talking	√	

3.3 Data base of room impulse responses acoustical background from different environments

Data were captured using an artificial human head and torso in an anechoic chamber and in different realistic environments: two offices, one cafeteria and one courtyard. Impulse responses were measured using in-ear microphones and behind-the-ear hearing aids. Additionally, the ambient sound in each environment was recorded.

This database enables the composition of natural acoustic scenes, allowing objects to be arbitrarily added to a scene by convolution of a clean mono recording of a desired object with the according impulse responses. Moreover, the signal-to-noise ratio of object to ambient sound is freely adjustable.

Detailed descriptions can be found in Deliverable D2.1: *Data Recorded from Different Acoustic Environments* and in a publication within the context of a special issue on "Digital Signal Processing for Hearing Instruments":

H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, *Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses*, EURASIP Journal on Advances in Signal Processing, vol. 2009, Article ID 298605, 10 pages, 2009.
The database is publicly available at http://medi.uni-oldenburg.de/hrir/.

3.4 Oldenburg logatome corpus (OLLO)

The OLLO database had been recorded under the EC project DIVINES (Diagnostic and Intrinsic Variabilities in Natural Speech, (FP6, STREP, 1st February 2004 - 31st January 2007) and has been used under DIRAC in the joint physiological and modelling work on biological and engineering approaches to acoustic detection. It is freely available under the address
http://sirius.physik.uni-oldenburg.de/html/download_ollo.html.

3.5 Neurophysiological Data Base

Most of the neurophysiological experiments carried out by partner LIN rely on stimuli that are parametrically well analyzed and are shared between laboratories by communicating the selected parameter spaces (e.g. pure tones, characterized by frequencies, durations, intensities, ramp functions, or rippled-noise stimuli characterized by ripple-modulation frequencies, etc.).

An exception might be the particular stimulus data based used for the envisaged odd-ball experiment to be conducted at the facilities of LIN as a collaboration between WP5 and WP2. This data base consists of a collection of frequency-modulated tone stimuli from which samples are taken according to predefined stimulus statistics for the different experimental phases. The corresponding stimulus sets are available from the DIRAC web page for internal use.

3.6 ETH/KUL locomotion stimuli

The stimuli were generated using motion-capture data from 6 male human adults (age ranging between 20 and 40 years) of average physical constitution that were walking or running at various speeds on a treadmill. Walking speeds were 2.5, 4.2 and 6 km/h while running speeds were 8, 10, and 12 km/h. The data were recorded at the Motion Capture Laboratory of ETHZ (Zürich) using an optical MoCap system (VICON) with 6 cameras operating at 120Hz and a spatial resolution of 1 cm. In order to reconstruct the 3D body motions, subjects wore a skin-tight suit with 41 infrared-reflective markers placed on the major anatomical landmarks. The

trajectories of the individual markers were then tracked and integrated into a 3D body representation. Based on these motion-capture coordinates, different displays were constructed using commercially available animation software (Maya, Autodesk Inc., USA) or Matlab (The Mathworks, USA) for each speed and actor. In the standard locomotion conditions, the displays consisted of humanoid figures where body limbs were represented by cylindrical geometrical primitives. From the 10 s long motion-captured movies (60 Hz frame rate), segments of 1000 ms or, in later sessions, 1086 ms were extracted, approximating one full walking cycle for the standard walking speed of 4.2 km/h. The starting positions of the locomotion cycle were varied across the different movies by sampling up to 109 different segments from the full 10 s movie.

The motion-captured locomotions were rendered at 8 different facing directions: 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315°. The 0°, 45°, 90°, 270° and 315° displays were generated based on the motion-captured 3D coordinates, while the other three remaining facing directions (135°, 180° and 225°) were obtained by mirroring the frames of the 45°, 0° and 315° displays, respectively. For each facing direction, the agent could move either forward or backward. Backward locomotion displays were created by reversing the temporal order of the frames of the forward locomotions. Thus, the snapshots of forward and backward locomotion displays were identical and differed only in their sequences. The stimuli employed in the electrophysiological studies consisted of approximately a full walking cycle and lasted 65 frames, equivalent to a stimulus duration of 1086 ms when played at a 60 Hz frame rate. Other control stimuli were made at KUL, including point-light displays, stick figures, upper and lower half of the body. At KUL, we have created also movies in which the walker suddenly reverses direction (from forward to backward walking or vice versa) to study at the single cell level the impact of a rare event on neural responses.

Some of the stimuli are available as a link to the J. Vision Vangeneugden et al. paper on the Journal of Vision website and all will be shortly available at the DIRAC website.

3.7 BUT

In April 2009, BUT started to collect a small amount of audio-only English speech recordings in low quality (8kHz, ogg compressed) for the demonstrator of the neural-net based OOV word detector. These recordings contained OOV words and unexpected sounds. So far, the data set comprises 16 small sets of utterances (1 female/6 male speakers, mostly non-native), corresponding word recognition output, neural net scores and phone posteriors. This set will be made available in a bundle

with the demo software on the DIRAC wiki page before the final WP6 deliverable is due.

In 2010, OHSU provided BUT with three recordings of an imaginary telephone call (each about 2 minutes of duration) covering a predefined list of OOV words (e.g. Barack Obama, mycorrhiza funghi, lithophytes orchid...).

Audio, the list of OOV words, and recognition output will be uploaded to the DIRAC wiki page.

Apart from that, BUT is using multiple speech data bases of annotated telephone and lecture speech to develop and evaluate the OOV detection in the scope of WP2 (see D2.12, section 1.4, table 1).

4 Validation Method

In order to validate the experiments on the data given by the data bases, the data has to be annotated first. This annotation has to be manually since no (semi-)automatic procedure is known to be 100% reliable.

This ground truth will serve as the reference against which the results of the test will be measured.

4.1 Annotation

The annotation of a data base item has to list actions / events which are to be interpreted as producing an incongruency when processed by the models of the different experiments. A suitable way is to use a time scale with cue words marking the begin of each event, combined with a more general description of the total scene. The Audio-visual recordings by FRA have already been annotated on a time scale accurate down to a second (see Appendix A for more details).

4.2 Measures

A score table per experiment will give the numbers for true and false positives as well as false and true negatives. This score table will be the aggregation of the experiment results on each test item (take, text etc.).

5 Validation Plan

In this section, a plan for the validation of the experiments and data already aggregated by the partners or planned will be presented. The subsections will address the selection of experiments and data, the collection of additional data, the data pre-processing, analysis and validation and a time schedule. The specific task for each validation step will be listed for each partner involved.

5.1 Selection of experiments and data

5.1.1 CTU

CTU will test audio-visual detection and reaction to incongruencies on data with multiple speakers, potentially in a dialog setting. In particular, it is important to test the distance of visual and audio detection for the speaker not facing the camera.

5.1.2 OL

OL will process all scenes that contain either defined places/movements of sound sources (localization) or defined activities of acoustic objects (classification) or both.

5.1.3 BUT

BUT will process the remaining OHSU OOV talk data and provide OOV detection results.
The newly recorded meeting data has to be annotated in order to perform Out-of-language detection and processed by our existing systems. Possible modifications to the system have to be taken into account for the changed task.

5.1.4 FRA

FRA will evaluate all relevant scenes of the audio-visual data base with the "conversation detector". Relevant scenes are scenes with one or more people in conversation / having a monologue.

5.1.5 ETHZ

ETHZ will validate all developed tracker tree approaches on different data in order to point out advantages and drawbacks of the existing techniques. The data is annotated manually, a per frame flag that indicates normal or abnormal is sufficient for validation.

5.2 Collection of additional data

5.2.1 OHSU

OHSU will record the following scenes requested by ETHZ and OL:

Scene 1:
Person walks around, talks normal for some time (head set microphone), then starts to mention a name several times (OOV word)

Scene 2:
Person walks around, sits down on the sofa, switches on the TV (should play some scene with speech!), switches off TV, does nothing (1-2 seconds), person starts talking, after some time mentions a name several times (OOV word)

Scene 3:
Person walks around, talks normal, picks up something (bag/book/shoes/...) from the ground, walks away

Scene 4:
Person walks around, talks normal, falls on the ground, calls for help (mentions name of a doctor/nurse)

Scene 5:
Person lies in bed, talks normal for some time, then calls for help (mentions name of a doctor/nurse)

Scene 6:
'Lunch/Dinner' scenario (1 or 2 people):

- P1 sits at a table eating (or at least pretending to).
- ~1 minute of lunch noises (e.g. cutlery on plates, chair rocks when person gets up to reach out for something at the far end of the table.
- during "dinner", P1 accidentally drops his cutlery to the floor (should be clearly audible). Alternatively drop the plate
- bends down to pick it up (acoustic rare event of drop sound, plus visual rare event of unusual movement/body position)

If two people are involved and they talk, each utterance has to be at least 4 seconds long.

Scene 7:
'Visit' scenario (2 people):

- P1 walks around,
- phone rings
- P1 answers, talks on the phone
- after ~10 seconds, P2 knocks at the door (from the outside ;)
- P1 ends the call, says 'come in'
- P2 enters
- short hellos

5.2.2 CTU, FRA

For the audio-visual detection and reaction to incongruencies on data with multiple speakers, potentially in a dialog setting, CTU will collaborate with FRA to acquire suitable test data.

5.2.3 FRA

FRA will assemble a catalogue of recordings of very basic, atomistic scenes with both the AWEAR-II and the OHSU setup under a very controlled environment. Very basic means for example only one person, only one action/incongruence, indoors, no cast shadows, enough light, defined sound sources. The scenes and the necessary restrictions will be defined with a high degree of detail. Thus, it will be easier to compose more complex scenes from these atomistic scenes without having too much uncertainty about what we can process with the detectors. These composed scenes again will be processed and evaluated with the different experiments.

5.2.4 OL

OL is performing ongoing recordings of isolated non-speech sounds for artificially constructing congruent and incongruent acoustic scenes. These data will facilitate performance evaluation of audio-only algorithms for classification and unexpected event detection. All recordings are labelled with the respective object classes.

5.2.5 BUT

BUT is planning to do recordings of 5 or more short meetings to evaluate the Out-of-language detection. The dominant language will be English, interrupted casually by other languages such as German or Czech.

Reference labels for the language of each segment in the meeting will be provided in addition.

5.3 Data pre-processing

5.3.1 CTU

CTU provides scripts and functions for low-level processing of the video stream, namely debayering and image rectification, to other partners.

5.3.2 BUT

The remaining audio recordings of OHSU will be processed through the hybrid word/sub-word OOV detection system. Results will be presented in terms of Hits and False alarms, in addition with the recognition output.

The newly recorded BUT meetings will be segmented automatically into speech/non-speech using a voice activity detection. Finally, the audio and the segmentation will be processed by our systems.

5.3.3 FRA, OL, ETHZ

FRA will pre-process the input data for the "conversation detector". The data from the three involved detectors, i.e. tracker tree, audio localization and speech/non-speech classification will be either provided from OL and ETHZ or will be processed by FRA using the detectors provided by the partners. The output data of the detectors will be pre-processed in order to have a common time basis (12 frames/second) and to map the pixel-based localisation information to the directivity resolution of the audio localisation.

FRA will assist the partners with the pre-processing of the data for the different experiments.

5.4 Data analysis

5.4.1 Data analysis by OL

Partner OL processes the audio stream and provides the results of the analysis to all partners via the wiki server. The representation of audio data is compliant with the format agreed upon in order to ensure hassle-free audio-visual analysis and integration of the data.

5.4.2 Data analysis by ETHZ

ETHZ will process the DIRAC data with the developed tracker tree techniques, which automatically generate visual abnormality labels for each frame. We expect adjustments to be necessary for the adaptation to different recording scenes (ensure functioning background subtraction, retrain models of normality, etc).

5.4.3 Data analysis by BUT

Reference language labels will be created manually for each talk. Next, Out-of-language detection will be performed similar to OOV word detection in terms of Hits/False Alarms of the detected OOL segments, but based on the sub-word output generated by the generic part of the recognizer. We expect, that the system has to be adjusted accordingly to perform reasonably well.

5.4.4 Data analysis by CTU

Partner CTU provides (i) image data stabilized w.r.t. the ground plane for the person detector and (ii) person detection results compliant with the format agreed upon in order to ensure hassle-free audio-visual analysis and integration of the data.

5.4.5 Data analysis by FRA

FRA will analyse the pre-processed data from the three detectors with the "conversation detector". This will include the annotation of the data where necessary and the analysis of the fused, pre-processed data input with the specific model of the "conversation detector".

5.5 Validation

5.5.1 IDIAP, ETHZ

IDIAP and ETHZ will validate the algorithm for detection of incongruent human actions and subsequent learning. Experiments will be designed so to evaluate performance with respect to (a) accuracy in detect incongruent actions, and (b) number of labelled samples necessary to learn the new detected action. As to our knowledge, there is no algorithm that could directly compare with ours, we will compare separately on the two points listed above with the current state of the art.

5.5.2 OL, CTU

Both classification and localization algorithms will be evaluated on all available databases. A particular emphasis will be put on analysing scenes with unexpected events such as unknown object classes. The audiovisual incongruence detection will be further pursued in cooperation with CTU, evaluating the audio and video algorithms on all scenes with matched and mismatched audio and visual signal sources.

5.5.3 CTU, ETHZ, FRA, OL

CTU will cross validate image data processing internally and the processing for audio-visual scene analysis in collaboration with ETHZ, FRA and OL depending on their requirements.

5.5.4 BUT

BUT will evaluate detection performance of OOV and OOL detection of the self-recorded data and self-created reference transcript. Hits and False Alarms, Precision and Recall will be reported.

5.5.5 FRA, OL, ETHZ

FRA will evaluate the results from the "conversation detector" experiment based on the annotated data and the detector output data from OL and ETHZ. True/false positives and negatives will be reported.

FRA will assist with the validation of other experiments on request of the partners.

5.5.6 ETHZ

ETHZ will validate the tracker trees on the recorded data for the detection of visual incongruencies in human behaviour.

5.6 Time Table

5.6.1 OL

The new scenes with acoustic rare events (see 5.2) will be evaluated as soon as they are available. New detectors will have to be set up and optimized for the task, which will be finished by end of September.

5.6.2 IDIAP-ETHZ

The expected timeline for the work is:

- July: Fabian Nater from ETHZ will visit IDIAP for theoretical work on the algorithm
- August: Theoretical framework fully developed
- September: first implementation and preliminary tests
- October: final implementation, experimental evaluation
- November: submission to CVPR of obtained results

5.6.3 CTU

CTU will cross validate visual and audio-visual data for scene analysis in collaboration with FRA and OL depending on their requirements of the partners.

5.6.4 BUT

The expected timeline for the work is as follows:

- August: meeting data will be recorded
- September: segmentation and annotation of recorded data will be created
- October: processing of data by the OOV detection system
- November: evaluation, some possibly necessary tweaking

5.6.5 FRA

The expected timeline for the work is:

- July 31: recording of atomistic scenes finished
- September 30: validation of "conversation detector" finished
- Ongoing: Help partners with data acquisition, exchange, publishing of Data/results

5.6.6 OHSU

OHSU will have recorded the requested scene by end of August.

5.6.7 ETHZ

ETHZ will process data as soon as it is available. Techniques will have to be adapted for the OHSU data, models have to be retrained, and this will be done by the end of September (one month after delivering the data). Developments and testing of new techniques and adaptation of existing techniques in order to have a more flexible model of normal human behaviour is ongoing (expected submission to CVPR in November).

6 Appendix A: Labelling of the audio-visual recordings

6.1 Location of the label in the data base

Each take has a label (called label.txt) that includes information about date, location etc., as well as a detailed description of the scene. It can be found under: Recording/Take/label.txt in each sub-directory of the wiki data base.

6.2 Structure of the label

Each label includes a header with information about the name of the recording and the scene, the date, the location, the used device, frame rate und a placeholder for comments.

This would be a typical label:

Recording: AggressionScenes
Scene : 0_AggressionScenes_verC_RobertStephan
Date : 2009-12-16
Location : house front next to HdH (FRA Oldenburg)
Equipment: AWEAR 2.0b (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 | The defensive person begins conversation;
0003 | 0005 | persons_2,persons_interact,shouting | The aggressor suddenly hits the other one in the stomach;
0005 | 0010 | persons_2,falling,fleeing,shouting | Person breaks down, aggressor flees;
0010 | 0017 | persons_1,limping | Person limps away;

6.3 List of key words used for description

Here, a list of the key words (cues) used to describe the content of the recorded scenes:

- Standing, sitting, lying, walking, running, limping, stumbling, hesitating, fleeing, falling

- backwards

- persons_1, persons_2, persons_3, persons_N, persons_interact

- car, bike

- speech, shouting, oov

- noise

- loudspeaker, headset

- asig_iv, asig_ov

- dropout

- overlapping