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: D 6.1
Application Scenario: Updated Version

Date of deliverable: 30.12.2006
Actual submission date: 30.06.2007

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **IDIAP Research Institute**

Revision [2]

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)	

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.)

D6.1 Application Scenario: Updated Version

IDIAP Research Institute (IDIAP)
Eidgenoessische Technische Hochschule Zuerich (ETHZ)
The Hebrew University of Jerusalem (HUJI)
Czech Technical University in Prague (CTU)
Carl von Ossietzky University Oldenburg (OL)
Leibniz Institut fuer Neurobiologie (LIN)
Katholieke Universiteit Leuven (KUL)
Oregon Health and Science University (OHSU)

Abstract:
Identifying, defining and developing appropriate audio-visual application scenarios was one of the key challenges of DIRAC during the first year of the project. Over the last 18 months, the definition of such an application scenario has been discussed and agreed upon by the members of the DIRAC consortium.

We identified a socially and economic important area as common focus, i.e. future care for elders, motivating the development of an audio-visual device that would be able to learn daily routines of its user, help him/her navigate the environment, identify unexpected (and/or potentially dangerous) situations, and issue alarms when needed. This general application area has been further focused on two particular, very important tasks, i.e.

1) Detection and identification of persons that do not belong to a trusted group or that exhibit abnormal or incongruous behavior.
2) Detection and description of unfamiliar environments and situations that require a change in the user's state or behavior.

These two tasks allow for exploring a whole range of important and previously rarely explored research issues, both in life sciences and in engineering. In addition, they are expected to yield a number of important spin-off problems and applications and will serve as a focus of DIRAC research.

Table of Content

1. Introduction

All DIRAC partners recognize that appropriate, meaningful and socially relevant application scenarios are necessary for driving collaborative research and for providing sufficient challenges to the theoretical and engineering research. In particular, non-trivial problems are necessary for identification and/or acquisition of appropriate databases and will be instrumental in developing appropriate test cases and demonstrations. Over the course of the first 18 months of the project, the definition of such an application scenario has been discussed and agreed upon by the members of the DIRAC consortium. Special consideration has been invested to account for the sometimes diverging needs of both audio/speech and vision processing.

DIRAC's declared goal is the design and development of an environment-adaptive autonomous artificial cognitive system that detects and identifies unexpected low prior probability events by probing for relevant cues, autonomously adapting to new and changing environments, and reliably discarding non-informative data from multiple information-seeking sensors. This problem represents a key challenge for understanding of natural cognitive systems, as well as for the design and application of artificial cognitive systems. The ability to detect such rare events in dynamic real-world environments, as described above, calls for dedicated efforts to understand the layout of the environment, detect the persons and objects therein, and analyze their actions and utterances and the context in which those are encountered. While both vision and speech analysis have made remarkable progress over recent years in relatively restricted scenarios, we believe that robust performance under less constrained conditions can only be achieved by a combination of the different modalities and of the different processing levels within those modalities.

Life science research forms an integral part of DIRAC's research activities, since in living cognitive systems (both animal and human), we have systems at hand that have already evolved impressive mechanisms for rare-event-processing in single (uni-sensory auditory or visual) and combined (multi-sensory) data streams. While the neurophysiologic mechanisms underlying these capabilities are currently not entirely known (neither in the uni-sensory nor in the multi-sensory case), several lines of research in DIRAC aim at both unraveling these mechanisms and making them available for application in artificial cognitive systems. The application scenario is chosen to define specific questions that can be addressed by neurophysiologic research and can be used as a framework in which engineering applications based on physiological motivation can be tested.

A comprehensive, mathematical definition of "rare" stimuli is treated in a separate document that will be made available at M21 on the DIRAC wiki bibliography web-page. Here, for the sake of completeness, we give only an intuitive description. The detection of "novel" or "rare" stimuli and the subsequent generation of the appropriate response, e.g., classification, is a fundamental property of any intelligent system. The cognitive system response to such stimuli has been the focus of extensive research in cognitive science (psychology), neuroscience and computer science. Most of this prior research has been focused on stimuli that have low probability and are, therefore, statistical "outliers" with respect to a distribution of observations used for training of the system. In contrast, the definition of rare stimuli or events in DIRAC is based on a combination of two components: (1) the deviation of the distribution of the observed features from those expected by the cognitive system's best guess at the stimulus class and (2) an estimate of the importance or utility of the consequences of the response. Both of these aspects rely on a combination of a feed forward feature distribution evaluation process in combination with a feedback cognitive loop that provides the context-and-class based expectation and context-based estimate of the importance (utility). For example, when a listener hears an utterance, he attempts to classify it using his present knowledge and to determine the best fitting class, e.g. a word. He then compares the phonemes of the best fitting word to the distribution of the perceived phonemes. If these

distributions diverge, and if the consequences of misunderstanding are sufficiently significant, the utterance will be treated as a new word worthy of learning.

Although the goal of the DIRAC scenarios is to foster research and test results, but not to develop a practical system, it is worth noting that the research carried out under the project does have practical potential. In particular, sticking to the specific field of care for the elderly, there are several uses for the intended guardian angel applications. Apart from the often cited functionalities like watching out for dangerous traffic or suspicious people that may pose a threat, we also see possibilities that are equally important. One of those is to compensate for problems that come with short-term memory loss in the early stages of Alzheimer's disease. Such patients could live independently for longer periods if assisted in appropriate ways. Users could speak to a system, telling about their intentions, like going to the bakery. The system could then detect a state of confusion, like the user losing track or standing around without a clear purpose (e.g. not waiting for traffic to pass, or not talking to other people). In such case the system can remind the user of his or her original intentions. Similar uses can be found indoor, like the classical problem of turning on the gas to make coffee, and then forgetting about it.

The rest of this document is organized as follows: section 2 gives an overall description of the application scenario, the care of elders, and the relevance of DIRAC's research to it. Section 3 describes in detail the application scenario and how it can be seen as two sub application scenarios: 'inside-home monitoring' (section 3.1) and 'outside-home intelligent walking aid' (section 3.2). For both scenarios, we explicitly list the targeted audio-visual events (section 3.1.1, section 3.2.1), the necessary vision and audio capabilities (section 3.1.2-3.1.3, section 3.2.2-3.2.3), the contributions from engineers and life sciences (section 3.1.4-3.1.5, section 3.2.4-3.2.5) and the datasets to be used (section 3.1.6, section 3.2.6). Section 3.3 describes a first implementation of the application scenario, to be carried out by month 24. Section 3.4 defines our criteria for success, and section 3.5 describes the ongoing collaborations between partners through a connecting matrix. Section 4 gives details on the hardware platform AWEAR that is being built for data acquisition. Finally, conclusions are given in section 5.

2. Application Scenario

Care for the aging is becoming rapidly one of the major issues for the health care locally and globally. The escalating proportion of individuals over 65 with chronic diseases and with declining cognitive functions confronts the caregivers with economical, medical, and social challenges on a global scale, and the only economically feasible solutions rely increasingly on sophisticated technology including monitoring and cognitive assistant devices. A successful development of such assistive devices is predicated on the timely solution of the problems addressed by the DIRAC project. In particular, effective assistance and intervention requires in-time recognition of abnormal situations and identification of the most appropriate intervention.

To provide for a maximum focus of all DIRAC research, we have identified two particular problems that are often encountered in the care of elderly, i.e.

1) Detection and identification of unfamiliar persons (i.e. observed persons looking/acting/sounding/speaking abnormally or not belonging to a trusted group).
2) Detection and identification of unfamiliar situations and environments (i.e. situations in which one's own current state or actions are not appropriate).

Detection and identification of unexpected persons can be achieved using stationary monitoring arrangements of sensors. To detect new unseen scenes, the device may be mounted on a wheel chair (or a similar mobile frame). Ultimately, we are aiming for a portable device that will assist its users in their daily routines, where we will need to address a

number of additional problems arising from the movements of optical and acoustic sensors in realistic noisy environments, among them the need for fast audio-visual identification of changes in the scene (a problem that is attracting the attention of the hearing aid industry), the need for accurate identification and description of unexpected words in speech (that is of interest for machine surveillance), the need for detailed analysis of complex scenes captured by moving visual sensors (that is required in a number of practical industrial applications), the need for learning and categorization from small samples, the need for continuous model adaptation, and the need for efficient fusion of information from various information sub-streams (problems which are of great interest to the machine learning community).

We aim to build such a combination by creating cognitive interactions and feedback loops, both within the different modalities and between them. In such interactions, low-level information from individual cues is integrated to arrive at an understanding of the higher-level context in which the sensor data has been perceived. This contextual information from the cognitive levels is then fed back to constrain the models which are applied to low-level processing, thus making it possible to restrict the search space, resolve ambiguities, compensate for missing data, correct errors, and last but not least, to increase acuity to the abnormal.

3. Detailed Application Scenario

In the following, we present a detailed description of the application scenarios we will work on in the scope of this project. The scenarios are grouped along two directions. The first focuses on *inside-home aid* for elderly persons with a static setup and assuming familiar scene layout and geometry. The second concentrates on developing capabilities for an *outside-home intelligent walking aid*, i.e. a mobile sensor setup to be used in both indoor and outdoor environments, with the capability to provide assistance in unfamiliar and potentially dangerous environments. This division will ensure that the developed techniques will be sufficiently general to be of practical use for real-world applications.

However, it is important to note that our work will be research-driven, not application driven. In other words, we take challenging real-world applications as an inspiration to address important audio-visual research problems. We are not aiming to deliver a complete and marketable solution to a particular problem (which would also not be feasible in the scope of this project).

3.1 Inside-Home Monitoring

For this scenario, the monitoring device will be static. Research will focus on observing the people present, both visually and acoustically, and on detecting important audio-visual events deriving from people and their behavior and utterances.

Still, capturing distant audio signals at a sufficient quality to accommodate speech processing is a very difficult problem. Thus, in the first stage we will extend the monitoring setup with body-worn microphones. In later stages of the project, those may be replaced by the device's static microphones (or microphone arrays), if feasible.

3.1.1 Targeted Audio-Visual Event

- Intruder detection, i.e. detection of unfamiliar persons entering the monitored room by their trajectories, motion patterns, produced sound, voice, or speech deviating from a learned model of normal behavior.
- Detection and resolution of audio-visual incongruencies. Examples could be (but are not restricted to) a monitored person being classified as male by visual cues, but

exhibiting a female voice pattern; a person visually classified as belonging to a trusted group (e.g. medical doctors), but employing atypical vocabulary; a person switching to a foreign language while addressing another person unfamiliar with that language, etc.

In order to detect those events, the following basic capabilities are needed from the vision and speech side.

3.1.2 Necessary Vision Capabilities

- Detecting all persons visible in the room and tracking their motion (trajectories and body pose).
- Learning models for normal trajectories and walking patterns (sequence of body poses).
- Recognizing specific actions based on the observed motion patterns (e.g. walking, running, etc.), body shapes, and trajectories.
- Comparing observed trajectories and motion patterns with the learned models and detecting irregularities.
- Continuous model refinement and updating.

3.1.3 Necessary Audio Capabilities

- Low-level processing for sound signal enhancement and separation of interfering sounds to enable high-level acoustic classification and speech recognition.
- Adaptation of sound signal enhancement to different classes of surrounding sounds.
- Detecting general classes of surrounding sounds that are indicative of the users present situation and actions (e.g. kitchen sounds, music from TV/radio, presence of many simultaneous voices).
- Detecting specific audio events that are indicative of normal or unexpected actions (e.g. phone ringing, possibly without being answered; sound of door opening)
- Detecting the presence of speech.
- Recognizing persons by their voice pattern.
- Recognizing and transcribing speech and identifying unexpected speech patterns.
- Continuous model refinement and updating.
- Detection of out of vocabulary words/utterances.

3.1.4 Contribution from Engineering

The vision part comprises contributions from WP1 (wide-angle-of-view imaging, and image processing, low-level visual features, body part detectors for unconstrained motion) and WP3 (person detection and tracking, body pose estimation, action recognition). The speech/audio part requires contributions from WP1 (acoustic features, acoustic, signal enhancement) and WP2 (audio classification, speech recognition). Combining the two will build on contributions from WP4 (automatic object detection and localization, feature selection for classification, incremental learning) and WP5 (audio-visual data fusion).

3.1.5 Contribution from Life Science

In different (multi-partner) workpackages, distinctive aspects of rare-event processing will be investigated and provided as input to the engineering groups. While in WP2 and WP3 life scientists and engineers will collaborate on unisensory (auditory and visual, resp.) rare-event processing, WP5 will directly investigate multi-sensory, audio-visual, integration and how rare-event processing benefits from integrating auditory and visual information streams. On a more abstract level, in WP4 life scientists and engineers will collaborate on potential classification strategies used by living and artificial cognitive systems. This research provides the substrate for (later) project phases where "high-level" representations of multi-sensory data will be exploited to respond meaningfully to rare events.

Specifically, WP2 focuses on classification of auditory events and processing of unexpected auditory events. As unexpectedness in auditory data streams may pertain to both entire auditory objects (such as sudden unexpected sounds, or unexpected words or phrases in speech) and isolated auditory features (such as unexpected spectral and/or temporal characteristics of sounds or unexpected intonation or prosody in speech), the life science research in this work package is currently focused on understanding how feature selectivity of neuronal responses (currently studied at the single unit level) contributes to the coding of a whole and its labeling as being unusual. Partners LIN and OL are collaborating on characterizing feature selectivity of neurons driven by auditory stimuli. Experimental studies will be performed by LIN; their results and identified mechanisms will then be taken up by OL in order to implement them in an "adaptive filter" design for artificial systems.

WP3 focuses on vision research. The main vision contributions in the above scenario are visual body pose estimation and action recognition. Both are based on work in WP3 informed by neurophysiologic studies at KUL investigating how those tasks are performed in the macaque brain. In particular, both the computational models and the neurophysiologic studies are based on the same training stimuli, recorded with a Motion Capture setup at ETH as part of WP3. While the envisaged test data in the application scenario will obviously not be suitable for controlled neurophysiologic experiments, the observed real-world performance of various computational models, developed with the help of neuroscience, will allow us to get a better understanding of those models and will suggest further neurophysiologic experiments.

In addition, the targeted application tasks will require learning of features for classification from various audio-visual inputs. WP 4 provides an analysis and testing of how high-level categorical representations of objects are formed and subserve rare-event detection and processing. The employed methods in WPs 2 and 3 will benefit from insights gained in psychophysical and combined brain imaging studies performed at HUJI and LIN, respectively, in WP4 and will also suggest additional experiments to be pursued there. In addition, the employed algorithms will be driven by results from HUJI's studies of incongruence detection by humans.

In WP5, life scientists and engineers will closely interact to study the role of multi-sensory interaction for rare-event processing. The focus will be on the mechanisms implemented in biological cognitive systems by which past experience in one modality is used to guide responses to non-classifiable, rare and unexpected, events in another modality. Here partner LIN will perform such behavioral experiments in combination with electrophysiological recording from rodent sensory cortices. The results will be used to formulate a theoretical model of multi-sensory fusion and rare-event processing by partners IDIAP and OHSU.

3.1.6 Datasets

We are aiming at testing out the developed techniques using realistic data captured by OHSU in an actual elder care environment. As such data capture is however a longer-term process and in order to facilitate fast experimental progress, first experiments will be based on audio-visual data that is similar to the targeted elderly-care application, but that will be recorded using the experimental audio-visual recording platform described in Section 4 or similar audio-visual setups.

Data acquisition

One serious problem with studying rare and unexpected events is that the events to be detected and classified are rare, making it difficult to evaluate the success of an approach. The acquisition of real-life data would, therefore, require extensive and expensive longitudinal studies. There are two solutions: (1) to focus on those application areas where data are already collected and (2) to simulate normal and rare events. Our current proposal is to focus on the latter approach. In particular, we propose to use younger adults, in some cases actors, to simulate situations and conditions that are rare but important in the care for elders.

When performing such an approach, it is however necessary to verify that the employed engineering components still perform reliably in normal situations, i.e. when no unexpected event is present. We will therefore also have to run baseline experiments applying the individual modules to additional data outside the limited corpus we can collect by using actors. In particular, audio-visual data available from different existing data sets will be used to develop, train, and verify the different components of our system.
The data to be used in the initial algorithm development and evaluation process will be collected at OHSU to simulate the audio/visual environment of an elder confronted by one of a large set of individuals, each represented by a short audio/visual clip. The volunteer participants in this data collection process will be recorded in two situations: (1) the first situation simulates in-person encounters. The actor will approach the simulated elder and utter a sequence comprising a number of prescribed of words and sentences; (2) the same participants will be asked to simulate the initial segment of videophone conversations with the elders. Both of these will be collected using either a high quality audio/video system or, for the sake of consistency the system described in section 4. The data collection will be performed at OHSU and the volunteers for this study will be recruited from several universities in Portland OR; we expect to recruit at least 100 volunteers, each recorded in 8-12 staged situations.

3.2 Outside-Home Intelligent Walking

This scenario is aimed at developing capabilities for an intelligent walking aid to be used outside the familiar home environment. This comprises both indoor and outdoor settings (e.g. when the device's user is inside a shop or walking on the street), with the common theme that the environment is a-priori unknown. This makes the recognition problem far more challenging for both vision and audio, since suitable algorithms have to adapt to changing visual and acoustic environments. On the other hand, such a scenario will lead to more generalizable approaches, as they cannot be based on simplifying assumptions that may not hold in practical applications.

3.2.1 Targeted audio-visual events

- Detection of unusual situations or environments, e.g. another person walking on a collision course, a person approaching the device's user and starting to speak, the acoustic scene properties suddenly changing, etc. Depending upon the type of situation, detection would be initiated either by omnidirectional sound or vision analysis and would prime the other cues with its results.
- Detection and resolution of audio-visual incongruence, as described above.

3.2.2 Necessary vision capabilities

- Dynamic scene analysis, including Structure-from-Motion.
- Self-localization in order to provide navigational aids.
- Recognition of generic settings (e.g. indoor/outdoor), as well as of familiar places or locations.
- Detecting and localizing important objects in the environment.
- Detecting and tracking people in the environment and recognizing their actions.
- Learning models of dynamic behaviors.
- Comparing observed trajectories and motion patterns with the learned models and detecting irregularities.
- Continuous model refinement and updating.

3.2.3 Necessary audio capabilities

- Audio classification, e.g. for detecting car engine or honking sounds and nearby speech.

- Adaptation of audio to different acoustic environments such as indoor vs. outdoor, or ambient sources vs. localized speaker.
- Directional acoustic filtering (e.g. beamforming) for enhancing sounds for certain direction and suppressing for others.
- Directional estimation of dominant acoustic source under benign acoustic environments.
- Speech recognition for indoor locations.

3.2.4 Contributions from Engineering

Implementation of the out-of-home scenario will build on low-level features developed in WP1 for both vision and audio. The vision part will make use of omnidirectional sensors for self-localization and automatic scene layout estimation using the bottom-up Structure-from-Motion and reconstruction framework developed in WP1 and WP3. This framework will be combined with top-down influences from object detection and tracking developed in WP3 and WP4 and will lead to the construction of cognitive loops. Detailed analysis of observed persons will be made possible by integrating body pose estimation and action recognition capabilities from WP3.

The audio part will use low-level modulation features from WP1 and statistical signal models from WP2 to apply audio classification methods from WP2 and speech recognition methods from WP2 and WP4. This classification information will be used in a feed-back loop to trigger and adapt low-level multi-channel signal enhancement developed in WP1, which in turn leads to more salient input for the audio classification and speech recognition algorithms. In low-reverberation or close-distance situations, multi-channel audio will be utilized for direction estimation of detected sources.

Audio and visual information streams will be combined with the help of contributions from WP4 (classification, feature selection, incremental learning) and WP5 (audio-visual data fusion).

3.2.5 Contributions from Life Sciences

Visual action recognition will again rely on insights from the close collaboration between neurophysiologic and computational studies performed in WP3. This concerns especially the visual body pose estimation task, for which we hope to be able to employ the same basic methodology as in Scenario 3.1, helped by the close integration of many different vision components in cognitive loops. As in the indoor Scenario 3.1, the recorded test data will not be suitable for controlled neurophysiologic experiments. However, the realistic test scenario will allow us to test out how various hypothesized computational models perform on non-trivial tasks in a challenging real-world environment, thus leading to a better understanding and refined models.

In addition, the audio classification task will draw inspirations from neurophysiologic studies of acoustical representations in WP2. Moreover, both visual and acoustic modalities will benefit from the results of psychophysical studies on human classification and incongruence detection from WP4. The combination of vision and audio, finally, will be guided by neurophysiologic studies on audio-visual fusion in WP5.

3.2.6 Datasets

Audio-visual data for this scenario will be captured using the mobile AWEAR hardware platform described in Section 4. Close integration of vision and audio components is especially important in this respect, as no other source of similar data has been available prior to the DIRAC project. For developing the individual components, we will make use of test recordings obtained by different prototype platforms until a first corpus of AWEAR data is available.

3.3 Initial Experiment Scenario

In the following, we present a detailed description of a first scenario implementation motivated by the above goals that will allow us to carry out a series of experiments which will help to build the AWEAR demonstrator and to assess our current abilities to integrate vision and audio processing. The goal is to construct a processing pipeline that will include contributions from all engineering partners and will generate problems to work around for other partners. We suggest a simple scenario which we can stage and process. We assume that the processing will be done off-line at different places, but the result will be a sequence of intermediate results and a data interface protocol. The result of the experiments will suggest the level of integration we can realistically aim at.

Our goal is to demonstrate results for a first (simple) version of this scenario at the Y2 review meeting. The scenario will then gradually be extended as new capabilities become available.

3.3.1 Base situation:

The user is standing in a room or walking in the street with the AWEAR setup. A person approaches and becomes visible in the omni-directional image. CTU will stage the outdoor scenario on the street under favorable lighting conditions. 180 degree field of view camera will be calibrated and tracked, images will be rectified to a box projection and geometrical constraints for pedestrian recognition will be computed.

Step 1:
This event is detected visually, the person is localized and tracked, and additional information is collected about his/her behavior with respect to the environment. Based on the current position estimate, the system selects the best perspective cutout view of the person for visual processing. (ETH, CTU, KUL-cv)

Interesting questions that could be addressed at this level include

- Is the person walking normally? (drunk people could be dangerous, running people could knock over an elderly person) (ETH, KUL-neuro)
- Does the person belong to a known visual group? (e.g. persons in police uniforms/white doctor cloaks/etc. can be trusted) (ETH)

Step 2:
The person approaches further and starts to speak. Vision localizes the person's face (relative to its body), records a close-up shot for gender identification (HUJI), and informs audio processing about its relative location for beamforming (OL) if permitted by ambient acoustics.

Interesting questions at this level

- Is the person male or female by its appearance? (HUJI)
- Is the person male or female by its voice? (OL)

Step 3:
Primed by first audio-visual analysis, speech/audio processing comes in and analyzes his/her utterances. (ETH, CTU, IDIAP, OL)

Interesting questions at this level

- Is the person speaking English or e.g. Czech? (addressing somebody in a foreign language is an unexpected event) (IDIAP)
- Is the person using unusual words (e.g. outside a very restricted ~60-words vocabulary, IDIAP) or speaking in an unusual manner (e.g. very slow or fast, OL)?

Step 4:
Cross-checks between the different modalities become possible (IDIAP, OHSU).

Interesting questions at this level concern incongruencies between the expectations of different sensing modalities. When those are detected, surprise will be the result, and a model update will become necessary. Example expectations include

- A person recognized as policeman/medical doctor would be expected to speak English.
- A person wearing certain clothing colors/styles would be expected to be a female.
- A person identified as male by visual gender recognition would be expected to also exhibit male speech/voice patterns.

Step 5:
The scenario allows us to dissociate low-level and high-level mechanisms of rare-event processing and ask the question of how both interact. Here, the physiological work on rare-event processing and multisensory fusion (LIN) provides a guiding input for the engineering work and theoretical modeling (IDIAP, OHSU).

- Low-level mechanisms of rare-event processing work in a bottom-up fashion and are based on the analysis of the stimulus input statistics into a physiological system (cf. P300, mismatch negativity, etc.). One focus of the physiological research will be the investigation of how the input statistics influence neuronal response properties. For these low-level mechanisms we will first use speech and speech-like stimuli.
- Another focus of physiological research are fast low-level (bottom-up) mechanisms of audio-visual integration which will be studied in human EEG. We expect that these low-level mechanisms can be put to use by the engineering groups more rapidly than high-level mechanisms, because the former might be possible to implement using available technology.
- High-level mechanisms of rare-event processing go beyond mere detection of a rare event (as evidenced by P300 or mismatch negativity) and involve processes towards specific appropriate responses of the physiological system. The physiological experiments by partner LIN, together with the theoretical analysis by partners IDIAP and OHSU, provide an experimental situation in which meaningful behavior to a rare visual event is made possible by high-level category transfer from the auditory modality to the visual modality (see WP5). This is an example of a high-level process (presumably implemented in neocortical activity) by which multisensory integration enables a system to produce appropriate *specific* behavior to a rare unexpected event.
- Although we do not expect this to happen within year 2, we also aim at implementing such high-level strategies in the application scenario in later project phases, when they are better understood due to our collaborative research in DIRAC. Nevertheless, our current activities provide the necessary building blocks for this endeavor.

3.4 Criteria for Success

By definition, recognizing audio-visual events requires the integration of both vision and audio components. In addition, the detection of abnormal or incongruous events necessitates powerful models of what is normal. In order to successfully address the challenging research tasks presented above, we therefore need to combine advanced capabilities in all areas. The individual system components will be developed in WPs 1-5, and the detailed evaluations presented there thus form important contributions also for the success of the later integration.

For the integration in the context of the application scenario, the detailed task specification from above allows to define quantitative performance metrics, such as *recognition rate* (what portion of the events are detected?), *accuracy* (how often is the reported result correct?), and *parsimony* (how much data do we need to see before an event can be detected?). Concrete

target values will depend on the actual test data and its complexity, but an important meta-criterion for the success of our integration is that it should provide a real benefit, i.e. that the integrated system using vision and audio cues will be able to detect audio-visual events which neither vision, nor audio can reliably detect on their own.

3.5 Connecting Matrix

A fundamental aspect of DIRAC research is its interdisciplinary, that requires strong collaborations between audio and vision research, and between scientists and engineers. As a way to promote and support such collaborations, we have introduced a "connecting matrix", i.e. a matrix showing which institute, and which researcher in each institute, is collaborating with, on a specific research issue. Because of the dynamic nature of work, this connecting matrix will evolve in time, and its most updated version will be reported every year in the updated Technical Annex. Here below we report the current version of the connecting matrix, with a detailed description of each ongoing collaboration.

	IDIAP	**ETH**	**HUJI**	**CTU**	**OL**	**LIN**	**KUL**	**OHSU**
IDIAP			X		X	X		
ETH			X	X	X		X	X
HUJI	X	X		X		X	X	
CTU		X	X		X		X	
OL	X	X		X		X	X	
LIN	X		X		X			X
KUL		X	X	X	X			
OHSU		X				X		

IDIAP-HUJI-OL

Partners are working together towards the implementation of an audio-visual object categorization algorithm able to recognize animals and objects like telephones or cars from both their visual appearance and their distinctive sound. HUJI has provided to IDIAP its object categorization software, and OL its sound recognition software. IDIAP is currently integrating these two pieces via an SVM-based fusion scheme that has proved successful for the integration of several visual cues. In the first step, experiments will be performed on static images, and we will investigate (a) the impact on performance of using multimodal information, with special attention to robustness, scalability and comparison with unimodal results; (b) detection of incongruities, like known visual input but unknown audio input (and vice versa), or known inputs but unknown combinations (e.g. a barking cat), as a concrete step towards detection of rare audio visual events. This will be contrasted with the psychophysical results on the detection of such incongruencies, collected in HUJI.

IDIAP-LIN

Partners LIN and IDIAP collaborate on the experimental/theoretical interface of rare-event processing for the tasks in WP5. Partner LIN performs the psychophysical and electrophysiological experiments with rodents revealing the neurophysiological mechanisms underlying the cross-modal transfer of information learned in one modality (audition) to interpret an other-wise unclassifiable event in the other modality (vision). These data are implemented in the general model of rare-event processing formulated by partners IDIAP and OHSU. The role of this interaction for DIRAC is that it provides an example (amenable to experimental and theoretical investigation) of how nature has evolved mechanisms by which audiovisual integration can be used for meaningful processing of rare events.

ETH-HUJI

Doron Feldman (HUJI) and Andreas Ess (ETH) have together performed experiments to test out the motion segmentation algorithm developed at HUJI together with the KUL-ETH

Structure-from-Motion system for the purpose of integration in a cognitive loop. In addition, collaborations in developing algorithms for object detection given small training samples are under way.

ETH-CTU

Several collaborations are in progress between CTU and ETH. Bastian Leibe (ETH) has run experiments with the ETH pedestrian detection approach on indoor surveillance data recorded by Tomas Svoboda (CTU) for the purpose of evaluating possible combinations with the CTU body part detection and unconstrained body pose estimation. In addition, CTU has recorded several omni-directional camera sequences for preparation of the AWEAR scenario. Hynek Bakstein and Michal Havlena (both CTU) have preprocessed this data and applied the CTU-KUL omni-directional Structure-from-Motion algorithm on it. The resulting data sets have been sent to ETH, where they are currently being processed with the ETH pedestrian detection and tracking system.

ETH-KUL

There have been several close collaborations between ETH and KUL, both on the vision and on the neurophysiology side. Bastian Leibe (ETH), Nico Cornelis and Kurt Cornelis (both KUL) closely collaborated on integrating Structure-form-Motion, object detection, and tracking in a cognitive loop for use in the AWEAR scenario. This collaboration has resulted in several joint publications (3DPVT'06, DAGM'06, CVPR'06) and scientific awards (CVPR'6 Best Video Award, CVPR'07 Best Paper Award). On the neurophysiology side, Tobias Jaeggli (ETH) collaborated with Rufin Vogels (KUL) on creating a novel stimulus set for use both in computational and neurophysiological action recognition. For this goal, human action sequences were recorded at ETH with a motion capture setup; the underlying motion data was extracted, analyzed statistically, and a computational body pose estimation and action recognition algorithm was trained on it. The same data set was then used to create several stimuli sets for neurophysiological experiments that are currently being performed at KUL.

ETH-OHSU

A collaboration has started between Tobias Jaeggly (ETH) and Misha Pavel (OHSU) to apply the ETH human body pose estimation algorithm to realistic OHSU data from a care-for-the-elderly scenario. A first test data set from OHSU has been sent to and processed at ETH, and code from ETH has been sent to and installed at OHSU.

OL-CTU-ETH

Partners CTU and OL are working together on audio-visual recognition of outdoor environments. CTU is applying its visual feature recognition algorithms to vision-based classification and OL contributes its audio classification algorithms which are based on a modulation analysis and an SVM classification stage. Cross-modal integration is performed by CTU based on integration of features and/or integration of the single-modal detection results. Initial work involves visual recognition of environments and audio recognition of presence of persons and cars. Later, we plan to combine it with visual recognition of cars and pedestrian developed by ETHZ. This work will eventually be benchmarked and merged with the work carried out by partners IDIAP-HUJI-OL on audio-visual object categorization.

LIN-HUJI

Partners LIN and HUJI collaborate on the experimental/theoretical interface of human categorization processes. Partner LIN provides the fMRI facility to combine the psychophysical experiment originally designed by partner HUJI on using positive vs negative equivalence constraints in visual categorization with brain imaging. Partners LIN and HUJI have collaborated to achieve fMRI compatibility of the experiment. The role of this interaction for DIRAC is that it will equip the consortium as a whole with the possibility to study brain mechanisms of auditory-visual interactions on a high processing level (categorization) in humans.

KUL-CTU
Jan Knopp from CTU is spending a three months research visit at KUL under the DIRAC training program, where he is collaborating with Mario Ausseloos and Kurt Cornelis (KUL) on performing self-localization based on the results coming out of the CTU-KUL omni-directional camera tracking algorithm. This will result in a new ability to recognize and localize oneself in a previously seen environment based on images.

HUJI-CTU
In order to deal with rare events, we would like to be able to learn from small sample as effectively as possible. One set of such algorithms, developed originally in HUJI, was designed to exploit training information which is given as pairwise constraints on pairs of data points. In parallel, CTU has developed a discrete optimization approach to the solution of segmentation problems with inter-point constraints. In our joint work, HUJI and CTU will formulate the problem of learning from equivalence constraints as a discrete optimization task for minimizing the number of contradictions, and design more efficient algorithms to solve the problem.

OL-KUL
The self-localization algorithm being developed by Jan Knopp (CTU), Mario Ausseloos and Kurt Cornelis (KUL) will be augmented with audio information to aid self-localization in previously seen environments where image and video data was pre-recorded. The integration of audio and imagery will improve robustness of the self-localization algorithm.

LIN-OHSU
Partners LIN and OHSU collaborate on the experimental/theoretical interface of (1) multisensory fusion and (2) rare event processing. Partner LIN performs physiological fusion experiments to dissociate low-level bottom-up from high-level top-down aspects of (1) and (2). Partner OHSU develops a formal theoretical account of (1) and (2) with particular respect to the question of at which level of organization (both in the sense of physical brain systems and abstract algorithmic organization) relevant sub-processes are realized.
The role of this interaction for DIRAC is the development (together with partner IDIAP) of a theoretical framework which (1) accounts for the relevant sub-processes of rare-event processing and multisensory fusion and (2) allows description of physiological, theoretical and applied aspects studied in DIRAC in a single uniform scheme.

HUJI-ETH-KUL
The motion segmentation algorithm of HUJI will be compared to a motion segmentation algorithm based on passive dense stereo reconstruction.

OL-LIN
Partners LIN and OL collaborate on the experimental/theoretical interface of processing rare auditory events. Partner LIN provides the measurements of STRFs of cortical receptive fields in rodents. This research reveals how preceding inputs to the cortex alter the selectivity profile (as estimated by STRFs) of neuronal responses. Hereby a link is established between the well-known field of "neuronal adaptation" (context-specific degradation of neuronal responsiveness to stimuli depending on their frequency of occurrence) and physiological characteristics of rare-event processing (like increased neuronal activity) to rare (or "odd") stimuli. Partner OL provides a theoretical treatment of rare-event processing in the auditory modality based on adaptive unsupervised learning paradigms. Repeated stimulus input from a stationary ensemble leads to an efficient (e.g. sparse) learned representation, which facilitates detection of novel stimuli in a superordinate hierarchical unit. Derived models will be applied to the classification tasks in DIRAC's application scenario.
The role of this interaction for DIRAC is that it provides, together with the physiological experiments and theoretical modelling in WP5, both a microscopic (single-neuron) and mesoscopic (neuronal assemblies) view of brain mechanisms of rare-event processing.

4 Hardware Platform (AWEAR)

In order to collect information-rich multimodal data from different audio and video sensors, we propose to build an experimental hardware platform. The construction of such a common recording platform is especially important for audio-visual integration, as available audio-visual data is very sparse and often restricted to static locations. While the construction of the platform itself is not the main goal of DIRAC, the construction of one or more prototypes will nevertheless be necessary in order to record audiovisual data and develop/evaluate our algorithms on it.

Some proposals for building a flexible hardware platform for audiovisual data collection (called AWEAR) have been made by KUL, CTU and OL and have been agreed upon by the consortium. This hardware platform is currently being constructed and will be funded from the special WP6 budget set aside for implementation of the DIRAC Application Scenario.

With the input of OL for the audio part and CTU for the vision part, KUL identified all hardware components and costs required in the first version of the AWEAR platform. To portray a rough picture of its content, we summarize the main elements of the AWEAR platform:

- 2 Firewire CCD cameras (omni-directional lenses).
- 4- or 6-channel hearing aid satellites and 2 high quality voice microphones
- 1 Firewire audio recording device with integrated pre-amps
- 1 or 2 laptops for recording the data
- 1 ultra-mobile PC or small internet tablet as control terminal

For quick development and easy handling, the initial version of the recording setup will be based on a cheap mobile platform (e.g. a child stroller). Depending upon our experience with this setup, it may later be extended with additional capabilities for its purpose of audio-visual data acquisition.

The choice of a behind-the-ears multi-channel microphone array as part of the audio sensors, which has been commented upon in the reviewers' report, appears to fit AWEAR's ultimate goal of being a fully-portable audio-visual aid, in particular for the elderly, very well. The sensor geometry comprises two near-linear microphone sub-arrays, one behind each ear, and combines properties of traditional linear arrays with human inspired engineering. E.g., standard beam-forming algorithms are available for each linear sub-array whereas the combination of both sub-arrays can in addition make use of inter-aural level differences (“shading effect” of the head) which psychophysics has shown to be highly important for human sound segregation and localization but which is unavailable as a cue for a single linear array.

A nearly identical sensor setup, albeit with less sophisticated signal processing, is already employed in the multi-channel hearing aids worn by many elders, hence there would be no additional burden of carrying the AWEAR audio sensors. Other forms of microphone arrays, such as larger linear or circular arrays, have not been accepted by end users at a large scale. The use of existing and widely deployed sensor technology clearly increases the potential impact of DIRAC's results, and considering the ongoing merge of technologies in the area of portable audio equipment (cell phones, head-sets, music players, hearing aids etc.) broadens the scope beyond AWEAR's initial application focus. One of the world's largest hearing aid manufacturers, Phonak AG, who’s chief technology officer is also a member of DIRAC's advisory board, has kindly agreed to provide us with the state-of-the-art high-quality microphone array. The additional cost for DIRAC using this sensor, as compared to one- or two-microphone solutions, is essentially zero

5 Conclusion

During the first 18 months of the project, the DIRAC consortium achieved the goal of identifying an appropriate audio-visual application scenario for a system able to detect rare audio-visual cues. Such a scenario is necessary for driving collaborative research among the partners and will be instrumental for developing appropriate demonstrations.

The identified scenario is centered around the theme of helping elderly persons to live in different environments with naturally occurring rare events, using an audio-visual cognitive aid. In particular, two important tasks, i.e. the detection and the identification of unknown persons, and the detection and identification of unknown environments, has been identified as prime applications areas that would aid in focusing the DIRAC research. Both of these tasks are of particular importance in home elderly care, where the ability of detecting deviations from normal routines of the patients is of great utility.