Inspérata accident magis saepe quam quae spéres.
(Things you do not expect happen more often that
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

FINAL ACTIVITY REPORT

2006-2010

Period covered from 01.01.2006 to 31.12.2010 Date of preparation: July 2011

Start of the project 01.01.2006 Duration 60 months

Revision [0]

Daphna Weinshall, **Hebrew University of Jerusalem**

Jörn Anemüller, **Carl von Ossietzky University Oldenburg**

Table of Contents

1. Executive Summary

The DIRAC project is an integrated project that has been carried out between January 1st 2006 and December 31st 2010. It has been funded by the European Commission within the Sixth Framework Research Programme (FP6) under contract number IST-027787. Ten partners from Belgium (KUL - Katholieke Universiteit Leuven), the Czech Republic (CTU - Czech Technical University in Prague, BUT - Brno Institute of Technology), Germany (OL - Carl von Ossietzky Universitaet Oldenburg, LIN - Leibniz Institut für Neurobiologie, FRA - Fraunhofer Institut Digitale Medientechnologie), Israel (HUJI - Hebrew University of Jerusalem), Switzerland (ETHZ - Eidgenoessische Tehnishe Hochschule Zürich, IDIAP - Fondation de l'Institut Dalle Molle d'Intelligence Artificielle Perceptive), and USA (Oregon Health and Science University) have investigated the concept of rare events in machine and cognitive systems, and developed multi-modal technology to identify such events and deal with them in audio-visual applications.

This document is the Final Activity Report of the DIRAC project, where we present the project and its achievements. In Section 2 we present **the research and engineering problem** that the project set out to tackle, and discuss why we believe that advance made on solving these problems will get us closer to **achieving the general objective** of building artificial cognitive system with cognitive capabilities. We describe **the approach** taken to solving the problem, detailing the theoretical framework we came up with. We further describe how the **inter-disciplinary** nature of our research and evidence collected from biological and cognitive systems gave us the necessary insights and support for the proposed approach. In Section 3 we describe our efforts towards system design that follow the principles identified in our theoretical investigation. In Section 4 we describe a variety of algorithms we have developed in the context of different applications, to implement the theoretical framework described in Section 2. In Section 5 we describe algorithmic progress on a variety of questions that concern the learning of those rare events as defined in our Section 2. Finally, in Section 6 we describe our application scenarios, an integrated test-bed developed to test our algorithms in an integrated way. Partner contribution is described in Section 8.

2. Rare events

The DIRAC project is about rare events. Why rare events? We motivate our question in Section 2.1. What are rare events? In Section 2.2 we propose a theoretical framework which answers this question. Finally, In Section 2.3 we discuss evidence from biological and cognitive systems, which support our choice of question and our proposed solution.

2.1. Motivation: the Problem

Since the introduction of von Neumann-like computing machines it was gradually becoming clear to most developers and users of information technologies that such machines, while increasing their computing power following the exponential Moore's law for many decades, are still failing behind biology on some seemingly very basic tasks. We have therefore assembled a group of likely-minded researchers, combining expertise in physiology of mammalian auditory and visual cortex and in audio/visual recognition engineering with the goal do discover what might be some of the fundamental issues that are preventing machines from being more effective on most cognitive tasks. All partners in the project agreed that understanding biological cognitive functions and emulating the selected ones in information extraction by machine is a way to achieving more efficient technology.

Among the fundamental machine weaknesses, we have identified one as particularly annoying: Machines work relatively well as long as the there is enough training data that describes well the information-carrying items that the machine needs to recognize. The machine fails when it encounters an unlikely or entirely unexpected item, typically recognizing it as one of the items from those that are expected. In contrast, it seems well founded that the unexpected items are the ones that get immediate attention when encountered by most biological systems. Addressing this fundamental discrepancy between the machine and the biological organisms is bound to produce some interesting challenges.

We therefore specified our aims, to design and develop an environment-adaptive autonomous active cognitive system that will detect and identify rare events from the information derived by multiple, active information-seeking sensors. The system should probe for relevant cues, should autonomously adapt to new and changing environments, and should reliably discard non-informative data.

Our first challenge was to define what we mean by rare events. One possibility was to focus on being able to recognize items that do not occur in the environment of a given sensor too often – this meant being able to deal with very small amounts of training data from the given modality. This is closely related to the problem of outlier detection. Alternatively, we may want to think about incongruity between modalities, or some incongruity between an event and the context in which it occurs.

During this first year of the project, we reached an agreement that the project is about detecting and identifying events that are unexpected from the system point of view, i.e. the events or items that have low prior probability given the previous experience of the system. We then came up with a principled way to go about this question, by comparing predictions made by several (at least two) systems with different degrees of prior experience.

In our final approach, described formally in Section 2.2, we have proposed and investigated a general, biologically-consistent, strategy to detect the unexpected low-prior probability events. Subsequently we built several applications that followed this principle. The strategy relies on

multiple information processing streams with different levels of prior constraints. This allows for the detection of instants where the incoming sensory data as evaluated by the more general model do not agree with the predictions implied by the more specific model. This strategy has been applied in detection of new faces for which a face classifier was not trained, new words that are not in the build-in dictionary of a speech recognizing machine and of new patterns of motion in video that do not obey constraints imposed by a model of expected human motions.

2.2. Definition and theory

Our definition of rare events is based on the observation that there are many different reasons why some stimuli could appear rare or novel. Here we focus on those unexpected events, which are defined by the incongruence between a prediction induced by prior experience (training data) and the evidence provided by the sensory data. To identify an item as incongruent, we use two parallel classifiers. One of them is strongly constrained by specific knowledge (either prior knowledge or data-derived during training), the other classifier is more general and less constrained. Both classifiers are assumed to yield class-posterior probabilities in response to a particular input signal. A sufficiently large discrepancy between posterior probabilities induced by input data in the two classifiers is taken as evidence that an item is incongruent.

Thus, in comparison with most existing work on novelty detection, one new and important characteristic of our approach is that we look for a level of description where the novel event is sufficiently probable. Rather than simply respond to an event which is rejected by all classifiers, which often requires no special attention (as in pure noise), we construct and exploit a hierarchy of representations. We attend to those events which are recognized (or accepted) at some more abstract levels of description in the hierarchy, while being rejected by the classifiers at the more specific levels.

More specifically, we assume that the set of labels (or concepts) represents the knowledge base about the stimuli domain, which is either given (by a teacher) or learned. In cognitive systems such knowledge is hardly ever a set; often, in fact, labels are given (or can be thought of) as a hierarchy. In general, a hierarchy can be represented by a directed graph, where each label (a set of objects) corresponds to a single node in the graph. A directed edge exists from label (concept) $\boldsymbol{a}$ to $\boldsymbol{b}$, iff $\boldsymbol{a}$ (the more specific concept) corresponds to a smaller set of events or objects in the world, which is contained in the set of events or objects corresponding to label $\boldsymbol{b}$, i.e., $\boldsymbol{a} \subset \boldsymbol{b}$. In this way the edges represent a partial order defined over the set of labels or concepts.

Because the graph is directed, it defines for each concept $\boldsymbol{a}$ two distinct sets of concepts (parent-child) related to it: *disjunctive concepts* which are smaller (subsets) according to the partial order, i.e. they are linked to node a by incoming edges converging on *a*; and *conjunctive concepts* which are larger (supersets) according to the partial order, i.e. they are linked to node $\boldsymbol{a}$

by outgoing edges diverging from ***a***. If the DAG of partial order is a tree, only one of these sets is non trivial (larger than 1).

We consider two possible tree-like hierarchies, which correspond to two intuitive cases:

Conjunctive Hierarchy: Modeling part membership, as in biological taxonomy or speech. For example, eyes, ears, and nose combine to form a head; head, legs and tail combine to form a dog; and sequences of phoneme constitute words and utterances. In this case, each node has a single parent and possibly many children.

Disjunctive Hierarchy: Modeling class membership, as in human categorization – where objects can be classified at different levels of generality, from sub-ordinate categories (most specific level), to basic level (intermediate level), to super-ordinate categories (most general level). For example, a Beagle (sub-ordinate category) is also a dog (basic level category), and it is also an animal (super-ordinate category), see right panel of Fig. 1. In this case, each node has a single child and possibly many parents.

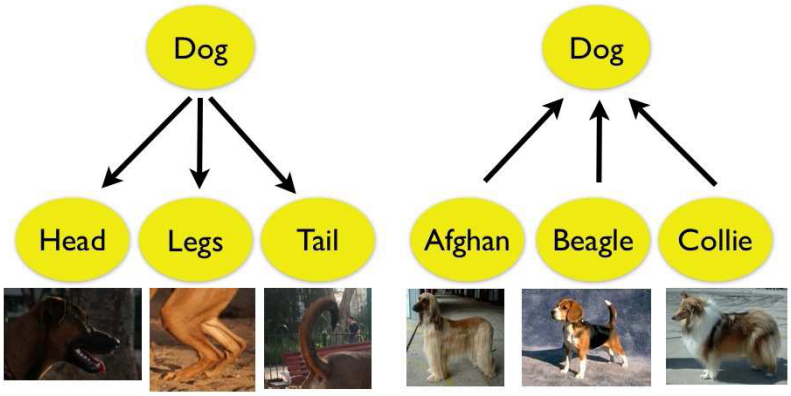

Figure 1. Examples. Left: Conjunctive hierarchy, the concept of a dog requires the conjunction of parts, including head, legs and tail. Right: Disjunctive hierarchy, the concept of a dog is defined as the disjunction of more specific concepts, including Afghan, Beagle and Collie.

Multiple probabilistic models for each concept

For each node ***a***, define A_s - the set of disjunctive concepts, corresponding to all nodes more specific (smaller) than a in accordance with the given partial order. Similarly, define A_g - the set of conjunctive concepts, corresponding to all nodes more general (larger) than a in accordance with the given partial order.

For each node a and training data T , we hypothesize 3 probabilistic models which are derived from T in different ways, in order to determine whether a new data point X can be described by concept a:

- $Q_a(X)$: a probabilistic model of class a, derived from training data T unconstrained by the partial order relations in the graph.

- $Q^s_a(X)$: a probabilistic model of class a which is based on the probability of concepts in A_s, assuming their independence of each other. Typically, the model incorporates a simple disjunctive relation between concepts in A_s.

- $Q^s_a(X)$: a probabilistic model of class a which is based on the probability of concepts in A_g, assuming their independence of each other. Here the model typically incorporates a simple conjunctive relation between concepts in A_g.

Definition of Incongruent Rare Events

In general, we expect the different models to provide roughly the same probabilistic estimate for the presence of concept ***a*** in data ***X***. A mismatch between the predictions of the different models may indicate that something new and interesting had been observed, unpredicted by the existing knowledge of the system. In particular, we are interested in the following discrepancy:

Definition: Observation ***X*** is incongruent if there exists a concept ***a*** such that $Q^s_a(X) >> Q_a(X)$ or $Q_a(X) >> Q^s_a(X)$. In other words, observation ***X*** is incongruent if a discrepancy exists between the inference of two classifiers, where the more general classifier is much more confident in the existence of the object than the more specific classifier.

Classifiers come in different form: they may accept or reject, they may generate a (possibly probabilistic) hypothesis, or they may choose an action. For binary classifiers that either accept or reject, the definition above implies one of two mutually exclusive cases: either the classifier based on the more general descriptions from level ***g*** accepts ***X*** while the direct classier rejects it, or the direct classifier accepts ***X*** while the classifier based on the more specific descriptions from level ***s*** rejects it. In either case, the concept receives high probability at some more general level (according to the partial order), but much lower probability when relying only on some more specific level.

2.3. Evidence from biological systems

In parallel to the theoretical investigation described above, we have investigated mechanisms underlying the detection of rare incongruous events in various ways. Thus, using a combination

of neurophysiological and behavioral experiments, we have rigorously demonstrated that the top-down mechanisms of rare-event processing we are interested in must not be confused with "mere novelty detection". In Section 5.3.we focus on a particular top-down mechanism, helping biological systems to respond meaningfully to unexpected rare events.

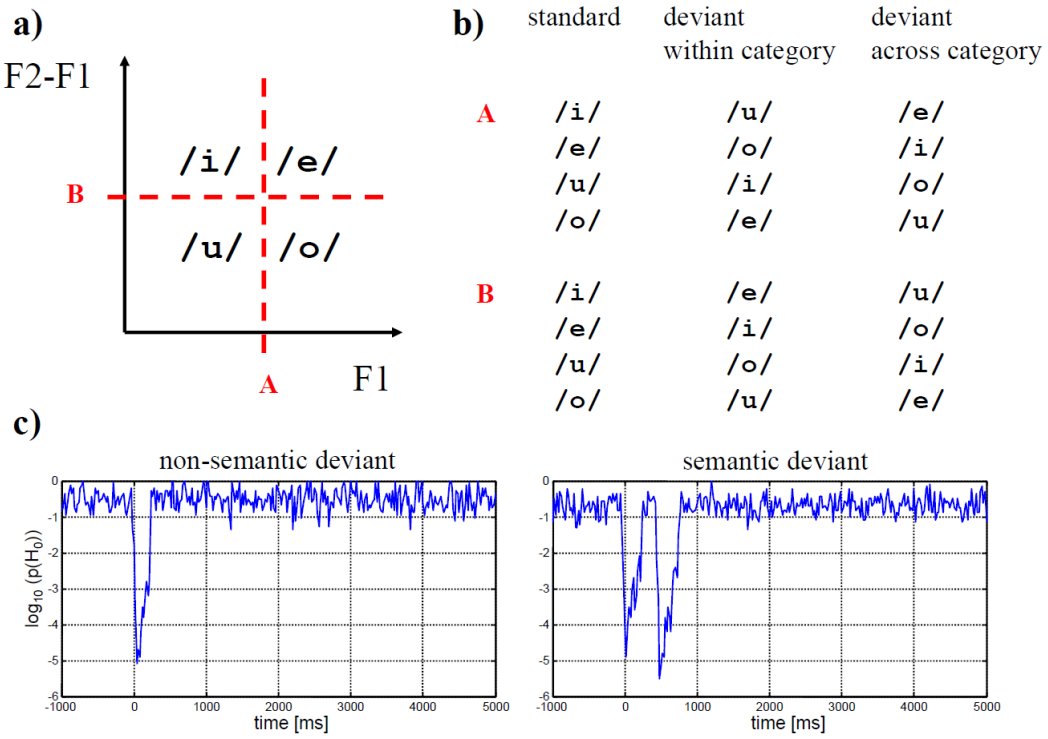

Figure 2. Panel a) shows the positions of the 4 vowels in the feature space spanned by the first formant (F1) and the spectral distance between the first and second formant. This space basically conforms to the classical vowel feature space descbribed by Peterson and Barney (1952), but was shown to be physiologically realized in mammalian auditory cortex (Ohl and Scheich, PNAS, 1997). One group of gerbils was trained to categorize the 4 vowels according to category boundary A, while the other group according to boundary B. The categorization training was realized as a standard active foot-shock avoidance Go-NoGo procedure, i.e. one category of stimuli required a Go response (jumping across a hurdle in a 2-compartment box) to avoid a foot-shock, the other category required a NoGo response (remaining in the current compartment of the box) to avoid the foot-shock. Thereby stimuli from the two categories acquire different meaning with respect to the appropriate behavior in the experiment. After training, classical novelty-detection ("odd-ball") experiments were conducted with both groups by presenting one vowel repeatedly as the standard stimulus and a second vowel as the infrequent deviant. Note that, given the previous vowel-categorization training, this second vowel could be selected either to be a member of the same meaning category as the standard stimulus or to be a member of the opposite category (associated with the opposite meaning with respect to required Go or NoGo behavior). Panel b) shows the different combinations of standard and deviants used in the post-categoriaztion tests. Note that in the two orthogonal classification schemes (A and B), physically identical stimuli played the role of within-category deviants and across-category deviants, respectively.

Traditionally, neuronal mechanisms underlying novelty detection are hypothesized to be reflected by the increased neuronal responses to deviant stimuli presented in the context of repeated, so called ”standard”, stimuli. This phenomenon is fundamentally a consequence of another well-known property of neuronal responses in sensory systems, namely stimulus-specific adaptation, i.e. the decaying neuronal response strengths with repeated presentation of identical stimuli. However, within DIRAC, we have emphasized that such bottom-up mechanisms of novelty detection can be conceptually and experimentally dissociated from top-down mechanisms of incongruence detection. Specifically, in one experiment we have been able to partial out the effects contributed by deviation of an unexpected stimulus with respect to its probability of occurrence (rareness) and contributed by deviating from the semantic context in which a stimulus can be expected. Moreover, the design of that experiment allowed us to make this comparison on the basis of neuronal responses to the exact same physical stimulus, which is a conceptual advantage to previous designs in this research area.

Specifically, in this experiment two groups of rodents (gerbils, Meriones unguiculatus) were trained to categorize 4 vowels from human speech in two orthogonally different ways. The two orthogonally different ways of forming a categorization boundary in the stimulus feature space allowed establishing two semantic contexts for identical features.

For neurophysiological analysis we recorded multichannel electrocorticograms from auditory cortex as these signals have been demonstrated to provide physiological correlates of category formation during learning (Ohl et al., 2001). Spatial patterns of electrocorticograms were used to classify vowel identity, and classification performance was analyzed in consecutive time bins of 120 ms (stepped in 20-ms steps) by comparing the number of correct classifications across all experimental trials with the expected number of correct classifications by chance (for details of the method see Deliano et al. 2009). For each empirically found number of correct classifications, panel c) shows the probability of observing this number of correct classifications by chance (null hypothesis), separately for deviants being a member of the same meaning class (non-semantic deviants) and for deviants being a member of the opposite meaning class (semantic deviants). Significantly ($p < 10\text{-}3$) different electrocorticogram patterns were found for both types of deviants at stimulus onset, but only for semantic deviants during an additional time window of 300 to 500 ms post stimulus onset. This latter result seems to indicate the existence of a physiological process mediating the detection of a top-down (meaning) incongruence, well separable from the bottom-up incongruence with respect to mere presentation statistics. More generally this phenomenon is in accordance with the general framework of incongruous-event detection in DIRAC as relying on detecting a mismatch between classifiers on a general level (vowel detectors) and a more specific level (vowels of class A or B).

A hierarchy of representations (e.g. from general to specific) is a fundamental assumption of the DIRAC framework. To test whether there is a hierarchical representation of animate and inanimate objects in the cortex, we have performed a monkey fMRI study in which we presented

images of monkey bodies, human bodies, monkey faces, human faces, man-made objects, fruits, sculptures, mammals and birds. We have also presented the same stimuli to humans in a human fMRI study. We have run 3 monkeys using both block and event-related designs. We have found patches of regions in the macaque Superior Temporal Sulcus (STS), coding for monkey bodies and animals (and for faces). The data provides evidence – at the fMRI macroscale – of a hierarchical representation of objects in the monkey and human brain. First, multivoxel pattern analysis of the activation patterns in two STS regions of interests showed that a greater selectivity in the anterior compared to the posterior STS. Second, the visual cortical regions in humans were activated by bodies of humans, monkeys, birds and mammals, while regions outside the visual cortex were specifically activated for images of human faces and human bodies, indicating again hierarchical processing: from general to specific, in agreement with the DIRAC framework.

3. System Design

In this chapter we describe our efforts towards system building, including work on visual sensors and the representation of visual data (Section 3.1), auditory tasks and the representation of auditory data (Section 3.2), and our own mobile system – AWEAR II (Section 3.3).

3.1. Visual sensors and representation

Inspired by human vision, we employed omni-directional cameras with 180 degree field of view. Image calibration and rectification based on calibration patterns as well as on matching images of unknown scenes have been implemented, tested and integrated into the processing of images from the mobile as well as static AWEAR platforms.

Feature extraction (such as SIFT, SURF, and MSER) has been tuned to omnidirectional images and extended to time-space domain in order to capture the geometry as well as motion of features in images and sequences. The first fully scale-invariant spatio-temporal feature detector that is fast enough for video processing or for obeying timing constraints, has been designed and successfully used for inter-video matching and action recognition.

Feature extraction and image matching have been verified by developing video processing platforms allowing to track cameras and localize the observer w.r.t. to an unknown scene. Thus, in combination with the new direction-of-sound-arrival detector, a new audio-visual sensor allowing to sense images and sound in a coherent observer-centered spatial relationship has been developed. The audio-visual sensor is able to deliver measurements in a static as well as in dynamic setup. The ability to register acoustic and visual sensing in a common (static as well as moving) coordinate system extended the state of the art in audio-visual processing and made it possible to detect incongruence between audio and video streams. It became the basis for incongruence detection demonstrations.

Motivated by far reaching applications in detection of rare events in surveillance and video processing, computational tracking and human action recognition schemes were developed and implemented to support incongruence detection for human actions at different semantic levels. HOG based human visual detection was gradually expanded by developing a number of specialized visual detectors for lower and upper human body parts and for human appearance correlated with different activities. The machine learning approach was used to extract relevant features of short activities and a method for visual-activity vocabulary construction has been designed. This was a key element for building tracker-trees for human activity detection and interpretation. Rather than being designed on the basis of pre-defined action classes, tracker-trees construction proved possible to automatically pick up the different types of actions present in the data, and to derive a tracker tree from those action classifications. It also demonstrated that it is possible to learn activities from few training data for those patterns classified as abnormal by the self-learned tracker tree. This makes it possible to let tracker trees evolve over time and react to changing environment and the appearance of new incongruence.

3.2. Auditory tasks and representation

Reliable detection and representation of acoustic events forms a building block in constructing audio- and audio-visual classifiers that extract meaningful information from the environment. Thereby, they serve to identify rare events and subsequently aid their further identification and, in the case of their re-occurrence, adaptive learning of new object classes.

Tasks to be accomplished for representing the acoustic environment can broadly be divided into two groups. Identification of the class as an acoustic event or object pertains to links a physical sound pressure wave to discrete object categories, such as “speech”, “car” or “dog”. In contrast, spatial representation of an acoustic scene typically requires at least two measurements of the acoustic wave–taken at different locations–that are processed in order to identify the location of acoustic events, largely independent of the category of the constituting events.

Categorization of acoustic events

Relevant categories for acoustic objects are context and task dependent. Scenes that contain at least some non-speech objects demand categorization of the sources’ identity, termed acoustic event detection. Speech recognition as a more specialized task necessitates the subdivision of the broad group of “speech” into categories relevant for ultimately recognizing an utterance’s meaning. Here, the about 40 speech phonemes are the appropriate categories.

Both tasks have been investigated in DIRAC using several sets of acoustic features that are extracted from the sound pressure waveform. A focus has been on robust features that are to a high degree invariant under environmental changes such as room acoustics and background noise,

leading to the analysis of the signals' modulation patterns, i.e., energy fluctuations across time in different spectral sub-bands.

The detection and discrimination of speech and non-speech objects builds on amplitude modulation spectrogram features (AMS). It represents a decomposition of the signal along the dimensions of acoustic frequency, modulation frequency and time, which is computed by a (modulation) spectral decomposition of sub-band spectral power time courses in overlapping temporal windows. The processing stages of the AMS computation are as follows. The signal decomposition with respect to acoustic frequency is computed by a short-term fast Fourier transformation (FFT with 32 ms Hann window, 4 ms shift, FFT length 256 samples, sampling rate 8 kHz), followed by squared magnitude computation, summation into rectangular, non-overlapping Bark bands and logarithmic amplitude com- pression. Within each spectral band, the modulation spectrum is obtained by applying another FFT (1000 ms Hann window, 500 ms shift, FFT length 250 samples) to the temporal tra- jectories of sub-band log energy. Outputs in the 0 Hz and 1 Hz modulation bands are influenced by DC components in the (log-energy) spectral domain and discarded as a means to reduce effects of channel noise (see also below). Finally, an envelope extraction and a further logarithmic compression are applied. By construction, the AMS features are approximately invariant to a time-domain signal convolution with short impulse re-sponses such as microphone transfer functions and early reverberation effects. Using these features, classifiers for specific acoustic objects such as "door", "keyboard", "telephone" and "speech" are learned using Gaussian-kernel support vector machines (SVM) using a 1-vs-all multi-class training approach. [Anemüller et al., Interspeech 2008; Bach et al., ICASSP 2010; Bach and Anemüller, Interspeech 2010].

Localization of acoustic events

The localization of sound sources in an acoustic scene is an important basis for the detection and identification of acoustic objects, albeit being somewhat orthogonal to the task of categorization since object category and object location are generally independent. Localization methods commonly employ more than one measurement channel, i.e., the spatial sound-field is sampled at several locations in space. Signals recorded at the different microphones differ in dependence on the room transfer functions from the sound sources to the different sensors, an effect that in a first approximation can be idealized as the time-delays with which the signal from a single source arrives at the different microphones. Cross-correlation measures and generalizations thereof (broadly termed "generalized cross-correlation", GCC) are employed to analyze these inter-channel differences. The ansatz developed here combines the phase-transform generalized cross-correlation measure (GCC-PHAT) with support vector classification. During training, a linear SVM is adapted to classify individual segments of the GCC-PHAT function as indicating presence (or absence) of a localized acoustic source. Thereby, a spatial map (parameterized by azimuth angle) is obtained that displays estimated directions of acoustic sources. The map shows the (possibly simultaneous) presence of sources, and in a post-processing

step source probability estimates for each time-point and each azimuth direction are computed from the SVM confidence scores.

The individual acoustic categorization and localization methods form modules that are combined in subsequent steps, outlined below, in order to construct full audio-only and audio-visual systems for the detection of rare events.

3.3. AWEAR 2.0 System: Omni-directional Audio-Visual Data Capture & Processing

To investigate the full scope of our approach, we designed an audio-visual backpack system which collects data that could be taken by smaller, wearable sensing systems. The primary goal by the Dirac partners was to look into the type of imagery which industry considers particularly important: cases in surveillance where cameras cannot be considered static (a problem with even fixed cameras on poles), traffic safety applications where the sensors are car- or pedestrian-borne, video analysis for automated summarization and retrieval, etc. Most methods assume static cameras, which often come with assumptions like foreground results from background subtraction, smooth motions, etc. As soon as such conditions break down, industry often finds itself without effective methods. Based on discussions with the reviewers, it was nonetheless decided that Dirac would focus on indoor applications, like the independent living one, where the AWEAR system still proved useful, but simply as a static apparatus.

(a) (b)

Figure 3. (a) The AWEAR 2.0 system is comprised of 3 computers, 2 cameras heading forwards, and a Firewire audio capturing device with 4 microphones (2 heading forwards and 2 backwards). Thanks to 4 lead-gel batteries, the autonomy is about 3 hours. (b) AutoCAD model of the system.

Sensor-wise, the AWEAR system is equipped with two high resolution cameras with fish-eye lenses, as well as a Firewire audio capturing device with four microphones, two heading forward and two backward. A total of three computers (two for video, one for audio, the latter also acting as the controller of the entire system) process the incoming multi-modal data streams, powered by a battery pack that can sustain the system for up to 3 hours. All components, along with further

supporting mechanic hardware, are mounted on a rigid frame. The total weight is 20kg (10kg of that for batteries). The system is shown in Fig.3.

The main benefits of the system are in particular (i) high resolution stereo, (ii) large field of view, and (iii) synchronization with multichannel high quality audio, and (iv) a wearable system. When searching for similar devices, one cannot find this combination.

Since not all the parameters were clearly known at design time, a modular design has been chosen. The platform can be extended to accommodate for four instead of two cameras or have the cameras replaced with faster ones capturing at double frame rate without having to modify the computing platform itself. Furthermore, up to four additional microphones can be added by just plugging them in. The computing platform has a margin in both bandwidth and processing power.

We used Ubuntu 8.10 as the operating system and several applications for video and audio capture. Video is captured in RAW (bayered) format into streams of 1000 files each, audio is saved as a 5-channel file, with the fifth channel containing the trigger pulses for video-audio synchronization.

For a system aimed at cognitive support, fish-eye lenses are very helpful due to their extended field of view. On the other hand, they required several dedicated steps for the data processing, going from calibration up to object class detection. Due to aberrations dependent on manufacturing and mounting, it is necessary to calibrate both lenses independently. For calibration, the entire field of view should be covered by a calibration target, rendering standard planar calibration targets unusable. We thus used a cube for that step. In order to find the transformation between the left and the right camera, we recorded a short sequence of 808 frames while walking in a room and then recovered the epipolar geometry as follows. Similarly, we developed debayering, geometrically rectifying, and several projection model cutout, structure-from-motion, image stabilization, and object class detection modules for AWEAR. For instance, to generate images more suitable for object recognition while keeping the full field of view, we used non-central cylindrical projection

4. The detection of rare events – algorithms

We designed a number of application-specific algorithms, which implement the theoretical framework described in Section 2.2 and adapt the theory to the specific application. Three application domains are described below.

4.1. Visual and audio object recognition

We adopted the framework described above to the problem of novel class detection, when given a **Disjunctive Hierarchy**. We assume a rich hierarchy, with non trivial (i.e. of size larger than 1) sets of *disjunctive concepts*, see right panel of Fig. 1. This assumption allows for the use of discriminative classifiers. We developed two applications: an algorithm to detect a new visual object, and an algorithm to detect a new auditory object.

Recall that in a disjunctive hierarchy we have two classifiers for each label or concept: the more general classifier $Q_{concept}$, and the specific disjunctive classifier $Q^s_{concept}$. The assumed classification scenario is multiclass, where several classes are already known.

Novel sub-classes of visual objects

In order to identify novel classes, our algorithm detects a discrepancy between $Q_{concept}$ and $Q^s_{concept}$. The classifier $Q_{concept}$ is trained in the usual way using all the examples of the object, while the specific classifier $Q^s_{concept}$ is trained to discriminatively distinguish between the concepts in the set of disjunctive concepts of the object. Our approach is general in the sense that it does not depend on the specifics of the underlying object class recognition algorithm. We tested the algorithm experimentally on two sets of visual objects: a facial data set where the problem is reduced to face verification, and the set of motorbikes from the Caltech256 bench-mark dataset.

Fig. 4 shows classification rates for the different types of test samples: Known - new samples from all known classes during the training phase; Unknown - samples from the unknown (novel) class which belong to the same General level as the Known classes but have been left out during training; Background - samples not belonging to the general level which were used as negative examples during the General level classifier training phase; and Unseen - samples of objects from classes not seen during the training phase, neither as positive nor as negative examples. The three possible types of classification are: Known - samples classified as belonging to one of the known classes; Unknown - samples classified as belonging to the unknown class; and Background - samples rejected by the General level classifier.

The results in Fig. 4 show the desired effects: each set of samples - Known, Unknown and Background, has the highest rate of correct classification in its own category. As desired, we also see similar recognition rates (or high acceptance rates) of the Known and Unknown classes by the general level classifier, indicating that both are regarded as similarly belonging to the same general level. Finally, samples from the Unseen set are rejected correctly by the general level classifier.

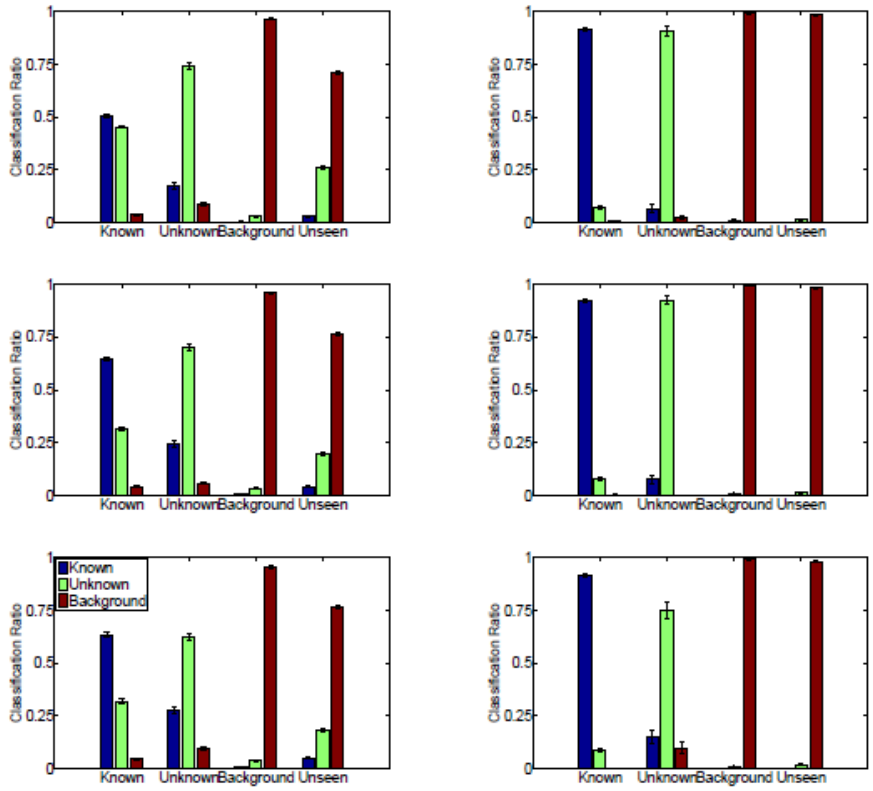

Figure 4. Classification ratios for 4 groups of samples: Known Classes, Unknown Class, Background and sample of unseen classes. Bars corresponding to the three possible classification rates are shown: left bar shows the known classification rate, middle bar shows the unknown classification rate, and right bar shows the background classification rate (rejection by the general level classifier). The panels on the left correspond to the Motorbikes general level class. The panels on the right are representative plots of the Faces general level class.

Novel sub-classes of auditory objects

We used the same algorithm as used above with an application from the domain of audio object classification, in order to evaluate the proposed framework in a different modality under systematically controlled noise levels. Here, the task is to discriminate known from novel audio objects appearing in an ambient sound background of a typical office environment. Hence, the inputs fall into three broad groups: Pure background noise (ambient environmental sounds such as ventilation noise recorded in an office room) with no specific audio object; known audio object embedded in background noise at a certain signal-to-noise ratio (SNR); and novel audio object embedded in the background at some SNR. Four classes of objects were considered: door opening and closing, keyboard typing, telephone ringing and speech. The non-speech sounds and the noise background were recorded on-site, speech was taken from the TIMIT database. The continuous

audio signals were cut into one second long frames, on which the analysis described below was carried out. Like before, performance is evaluated in a leave-one-out procedure, i.e., each of the office objects is defined as novel once and left out of the training set.

The resulting performance levels at equal error rate (EER) are displayed in Fig. 5. Here, the performance is bounded from above by the (arbitrary) choice of 5% false positive for the tuning of the general classifier. The results demonstrate that the detection of Unknown objects based on a hierarchy of classifiers is possible in the acoustic domain and its performance depends on the type of novel signal and SNR.

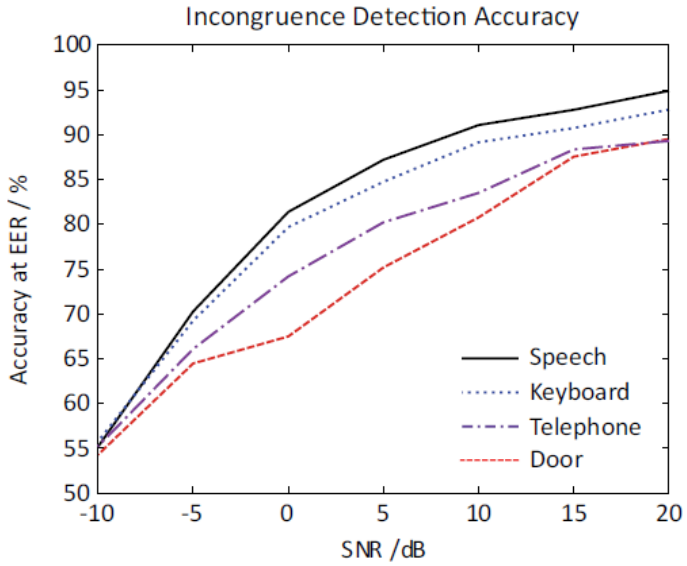

Figure 5. Accuracy of novelty detection, with one curve per type of novel audio object (see legend). The accuracy is taken at the EER point (equal false alarm and miss rates). Below 10dB, the EER could not be determined. Note that the accuracy is bounded from above by the (arbitrary) choice of 5% false positive rate for the general classifier.

4.2. Out of vocabulary words in speech processing

Current large vocabulary continuous speech recognition systems (LVCSR) are customized to operate with a limited vocabulary on a restricted domain. As prior knowledge, text-derived language models (LM) and pronunciation lexicons are utilized, and are designed to cover the most frequent words and multi-grams. Under real conditions, however, human speech can contain an unlimited amount proper names, foreign and invented words. Thus, unexpected lexical items are unavoidable.

If a word is missing in the dictionary (out-of-vocabulary - OOV), the probability of any word sequence containing this word according to such LM is zero. As a consequence, the

corresponding speech will be mis-recognized - the OOVs are replaced by acoustically similar in-vocabulary (IV) words. The information contained in the OOVs is lost and cannot be recovered in later processing stages. Due to the contextual nature of the LM, also the surrounding words tend to be wrong. Since OOVs are rare, they usually do not have a large impact on the Word Error Rate (WER). However, information theory tells us that rare and unexpected events are likely to be information rich. Improving the machine ability to handle unexpected words would considerably increase the utility of speech recognition technology.

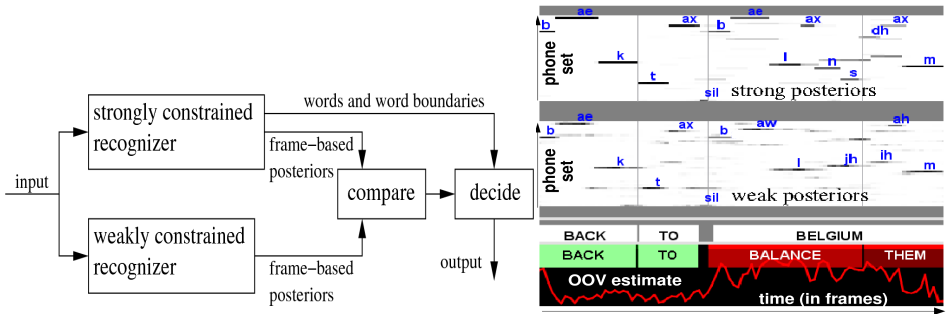

Figure 6. Left: Application of the incongruence model for OOV detection. The generic model is the weakly constrained recognizer and the specific model is the strongly constrained recognizer. Right: Phone posteriors and OOV detection output using NN-based comparison for the OOV “Belgium".

Within the DIRAC project a novel technique for the detection of unexpected and rare words (especially OOV) in speech has been proposed and developed. The approach is based on the comparison of two phoneme posterior streams (Fig. 6) derived from the identical acoustic evidence while using two different sets of prior constraints - strongly constrained (LVCSR, word-based, with LM) and weakly constrained (only phones). We aim to detect both where the recognizer is unsure and where the recognizer is sure about the wrong thing. The mismatch between the two posterior streams can indicate an OOV, although the LVCSR itself is quite sure of its output.

The hierarchical rare events detection scheme has been shown to outperform related existing posterior based confidence measures (CM) when evaluated on a small vocabulary task, where the posterior estimates in the two channels were compared by evaluating the Kullback-Leibler divergence at each frame. After that, the use of frame-based, word- and phone- posterior probabilities (“posteriors”) as CM was further investigated on a large vocabulary task (Wall Street Journal data) with reduced recognition vocabulary. The task was to classify each recognized word as either being OOV or IV, based on a word confidence score, obtained from averaging frame level CMs over the boundaries in the recognition output. With the introduction of a trained comparison of posterior streams (using a neural net – NN) and using new hierarchical

techniques for estimating phoneme posteriors, significant improvement over state-of-the-art posterior-based CM was achieved.

After that, the NN based OOV word detection was applied to noisy, lower quality telephone speech (CallHome, Eval01, Fisher) to show the robustness of the approach. In addition, the classification performance improved by classifying the recognized word output using several classes (IV correct, IV incorrect, OOV, silence).

The vocabulary used in speech databases usually consists of two types of words: firstly, a limited set of common words, shared across multiple documents (typically IV); secondly, a virtually unlimited set of rare words, which might only appear a few times in particular documents (mostly OOV). Even if these words do not have a big impact on the word error rate, they usually carry important information. OOVs which occur repeatedly often represent topic-specific terminology and named entities of a domain. Therefore, we further concentrated on the detection of frequently re-occurring unexpected words. To achieve this, we ran our OOV detection system on telephone calls and lectures centered around a certain topic.

For all further experiments, OOV detection should serve as an instrument to obtain descriptions of OOV words - putting special emphasis on repeatedly occurring OOVs. The task is now to detect the time span of the reference OOV word as completely and precisely as possible. Since the NN-based OOV detection system offers no description/boundaries of the OOV, we subsequently integrated a hybrid word/sub-word recognizer into our system (Fig. 7), which lets us obtain boundaries and descriptions of OOVs words in an integrated way. The system is not required to substitute an OOV by some IV; it can fall back to its sub-word model and thus retrieve a lower-level description of the word in terms of sub-word units.. Here the boundaries for OOV words were estimated more accurately than with the NN-based system.

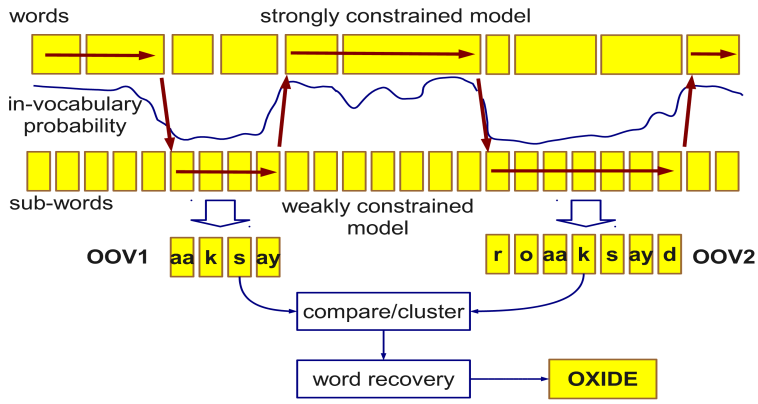

Figure 7: Schema of a hybrid word/sub-word model for OOV detection. The decoder has the freedom to compose the path (recognized word sequence) of words from either the specific word or the generic sub-word model. The detected sub-word sequences from the best path are taken as OOV candidates for similarity scoring and recovery.

The artificially high OOV rates due to the use of small decoding vocabularies were identified as a problem. Therefore, we finally changed to a much larger decoding vocabulary and a data set consisting of a collection of topic-specific TED talks, in which we find a reasonably high number of information-rich OOV words. It represents a more realistic scenario, since only naturally appearing and harder to detect OOVs are targeted.

To deal with rare and new words in ASR we proposed follow-up actions that can be taken after the detection of an OOV to analyze the newly discovered words and to recover from the mis-recognitions. The goal is to avoid mis-recognitions in the presence of rare words by designing a system that is open-vocabulary and that can learn with its usage. The hybrid word/sub-word recognizer solves the OOV localization and obtains its phonetic description – the detected phoneme sequence in the detected time span - in an integrated way.

Given the location and phonetic description of an OOV, one possible action is to recover the orthographic spelling of the OOV. We showed that OOV spelling recovery can successfully recover many OOVs, lowering the word error rate and reducing the number of false OOV detections.

Aiming at topic-specific repeating OOVs, we introduced the task of similarity scoring and clustering of detected OOVs. A similarity measure based on aligning the detected sub-word sequences was developed, which serves to identify similar candidates among all OOV detections. A new form of word alignment is introduced, based on aligning the OOV to sequences of IVs/other OOVs, which retrieves a higher-level description of the OOV, in the sense of word relations. (e.g. being a compounded word or a derivation of a known word).

4.3. Biological motion

In our third application, we developed a system for the detection of incongruent events that is based on the detection of activities, i.e. motion patterns. We compiled a set of motion data that contains walking and running patterns by several subjects, at different speeds. Subjects were placed on a conveyor belt, so that a motion capture system could capture the data. The system delivered the 3D coordinates of the body joints as they evolved over time. Moreover, normal cameras were also taking videos of the same actions, from which a series of silhouettes were obtained from 8 different viewpoints. The computational work then started from these silhouettes, whereas the neurophysiological work (described in the next section) started from different types of stick like figures, based on the motion captured data of exactly the same actions.

Of course, in order to tackle real-life problems, it was necessary to capture quite a broader range of activities. Thus we developed the so-called tracker trees. These are hierarchies of trackers, with a very generic blob tracker at the root node, and becoming more and more specific when moving to higher layers. For instance, a walking tracker has been trained, which would respond to the

walking pattern of any subject. At the higher level, a layer of individual walking trackers is found, which respond to the particular gait of the corresponding people. In this way, the tracker tree can spot incongruent events of very different nature, e.g. a dog entering the apartment if there normally is none, a person falling, or another person than those known to the system entering the house. The system can also indicate what kind of incongruency occurred, at least qualitatively. This is important as the kind of action to take depends on the nature of the particular incongruency.

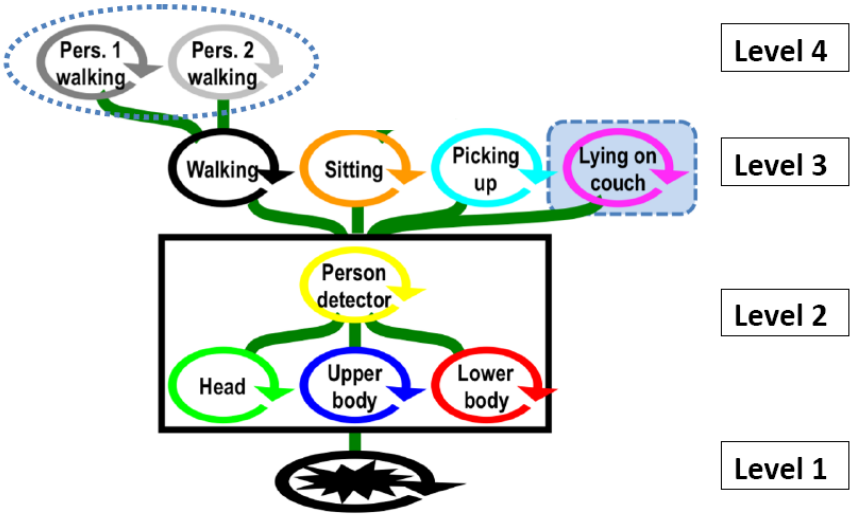

Figure 8: Visualization of the tracker tree and its dependencies over multiple levels.

We subsequently added several upper body activities detectors to the basic tracker tree, including such actions as reading or drinking. These required a more involved analysis than the rest of the activities, as they are more difficult to distinguish from silhouettes. We used our own spatio-temporal features - 3D extensions of SURF features.

In a further move, we extended the tracker trees with self-learning capabilities. This is important as in practical situations extensive training may be overly expensive, e.g. if one would want to install such systems in the homes of many people. First, a dual hierarchy has been developed, to mirror the snapshot - motion duality found in the neurophysiological research described in the next section. This type of tracker tree would automatically pick up new types of silhouettes or silhouette sequences. The response of this tracker tree has also been compared to neuronal responses to exactly the same dataset (see first paragraph). Strong qualitative similarities could be found.

We then set out to further improve the results and the flexibility of the self-learning tracker trees. A novel approach was introduced, where the hierarchy is discovered through Slow Feature

Analysis. Once activity nodes have been built, these are described on the basis of PCA, to well capture the variability among activities of the same class. This type of self-learning tracker tree allows for other than fixed splits (e.g. binary), as was assumed with the initial type.

In parallel we invested resources in exploring applications other than independent living, especially during the final phase of the project. Several extensions have been demonstrated, including outdoor usage, surveillance application, automated detection of abnormal images captured by webcams, etc.

Moreover, in order to allow for self-adaptive tracker trees and automatically add extra nodes, we integrated it with the transfer learning method described below in Section 5.2. After the initial tracker tree has been trained, it would start to detect events that according to the training material are incongruent. Applying the Transfer Learning principles we could automatically add new nodes for the incongruent events. This then allows us to name them, and take appropriate action for each case. In this way a small number of training events sufficed to build the additional nodes.

4.4. Biological motion as perceived by biological systems

In order to detect an incongruent event it is essential that one has a model of congruent events. Thus the aim of this section is to understand how biological organisms, in particular primates, represent actions of people. The computational principles discovered in these neurophysiological studies suggested and supported the computational framework described above in Section 4.3 and were used to detect actions, in particular walking/running people.

In a first series of studies, we studied the spiking responses of single macaque temporal cortical (rostral Superior Temporal Sulcus (STS)) neurons to a parameterized set of dynamic visual images of actions. We used arm actions like knocking, lifting and throwing and their morphs. The action images were rendered as stick figures. We found that as a population, the neuronal population represented the similarity among the different actions, as shown by a non-linear multidimensional scaling (ISOMAP) of the pairwise differences between the neural responses to the different stimuli. We were able to distinguish different kinds of neuronal selectivity. Firstly, neurons, mainly in the ventral bank of the rostral STS, responded as well to the action movies as to static snapshots of these movies. These neurons clearly responded to form information. Secondly, other neurons, mainly in the dorsal bank of the rostral STS, responded much less to static snapshots than to the action movies, thus responding to motion information. This dual processing, form and motion based, has been used to develop the computational framework for action detection described above in Section 4.3.

In the next series of studies, we employed stimuli that were based on motion-capture data of real human subjects that were walking or running at different, controlled speeds on a treadmill, as described above in Section 4.3.. In a first phase, we performed an extensive behavioral study of

the perception of these biological motion displays in monkeys. We trained 3 macaques in the discrimination of facing-direction (left versus right) and forward versus backward walking using the above discussed motion-capture-based locomotion displays in which the body features were represented by cylinder-like primitives. Discriminating forward versus backward locomotion requires motion information while the facing-direction/view task can be solved using motion and/or form. All monkeys required lengthy training to learn the forward-backward task, while the view task was learned more quickly. Once acquired, the discriminations were specific to walking and stimulus format but generalized across actors. Performance in the forward-backward task was highly susceptible to degradations of spatio-temporal stimulus coherence and motion information. Importantly, collaborative computational work showed that the walking-running speed generalization in the forward-backward discrimination fitted the predictions made using the DIRAC computational architecture developed by ETH, thus supporting this architecture.

After the behavioral training, we conducted single cell recordings in the trained animals, examining the contribution of motion and form information to the selectivity for locomotion actions. We recorded in both dorsal and ventral banks of the rostral STS. The majority of the neurons were selective for facing direction, while a minority distinguished forward from backward walking. We employed Support Vector Machines classifiers to assess how well the population of recorded neurons could classify the different walking directions and forward from backward walking. Support vector machines using the temporal cortical population responses as input classified facing direction well, but forward and backward walking less so but still significantly better than chance. Classification performance for forward versus backward walking improved markedly when the within-action response modulation was considered, reflecting differences in momentary body poses within the locomotion sequences. Analysis of the responses to walking sequences wherein the start frame was varied across trials showed that some neurons also carried a snapshot sequence signal. Such sequence information was present in neurons that responded to static snapshot presentations and in neurons that required motion. In summary, our data suggest that most STS neurons predominantly signal momentary pose. In addition, some of these temporal cortical neurons, including those responding to static pose, are sensitive to pose sequence, which can contribute to the signaling of learned action sequences. Both mechanisms, the pose mechanism and the pose-sequence mechanism, have been incorporated into the model described above in Section 4.3.

5. The learning of rare events

Now that we have detected those rare events, comes the question of what to do with them: How do we bootstrap some representation, or classifier, for these novel and yet interesting events? This chapter describe various studies which address related questions, such as online learning (Section 5.1) and knowledge transfer seen from a computational (Section 5.2) and biological (section 5.3) points of view.

5.1. Online learning

The capability to learn continuously over time, taking advantage of experience and adapting to changing situations and stimuli is a crucial component of autonomous cognitive systems. This is even more important when dealing with rare events, where learning must start from little available data. From an algorithmic point of view, this means developing algorithms able to build novel representations from a few labeled samples, which are made available one at the time. At the same time, it is desirable for the same algorithm to be able to cope with little incoming data (this is the case when learning a detected rare event), and be able to choose to update the representation of a known class for which large amounts of data are available. This calls for online learning algorithms able to cope with small amount of data as well as large amounts, without suffering from memory explosion, while maintaining high performance and low algorithm complexity.

In one of our attempts to address this issue, we developed an algorithm - Online Incremental SVM, where we proposed an online framework with a theoretically bounded size for the solution. The basic idea was to project the new incoming data on the space of the current solution, and to add it to the solution only if it was linearly independent. The drawback of the method was that all the incoming data had to be stored in order to have an exact solution. We applied the same principle of projecting the incoming data on the space of the current solution to the perceptron, an online algorithm with forgetting properties. The resulting method, which we called projectron], has bounded memory growth and low algorithmic complexity. Perceptron-like algorithms are known to provide lower performance than SVM-based method.

We then moved to extend these results to multi modal data. We developed a Multi Kernel Learning algorithm that is state of the art in terms of speed during training and test, and in terms of performance on several benchmark database. We have called the algorithm OBSCURE (for "Online-Batch Strongly Convex mUlti keRnel lEarning algorithm"). It has guaranteed fast convergence rate to the optimal solution. OBSCURE has training time that depends linearly on the number of training examples, with a convergence rate sub-linear in the number of features/kernels used. At the same time, it achieves state-of-the-art performance on standard benchmark databases. The algorithm is based on a stochastic sub-gradient descent algorithm in the primal objective formulation. Minimizing the primal objective function directly results in a convergence rate that is faster and provable, rather than optimizing the dual objective. Furthermore, we show that by optimizing the primal objective function directly, we can stop the algorithm after a few iterations, while still retaining a performance close to the optimal one.

5.2. Knowledge transfer

How to exploit prior knowledge to learn a new concept when having only few data samples is a crucial component for being able to react upon the detection of an incongruent event. This can

be declined in two settings: transfer learning across models build on the same input modalities, and transfer learning across models built over different modalities. In both cases, we addressed the problem of transfer learning within a discriminative framework.

When dealing with knowledge transfer across models built on the same modality, the basic intuition is that, if a system has already learned N categories, learning the N+1 should be easier, even from one or few training samples. We focused on three key issues for knowledge transfer: how to transfer, what to transfer and when to transfer. We proposed a discriminative method based on Least Square SVM (LS-SVM) (how to transfer) that learns the new class through adaptation. We define the prior knowledge as the hyperplanes of the classifiers of the N classes already learned (what to transfer). Hence knowledge transfer is equivalent to constraining the hyperplane of the N+1 new category to be close to those of a sub-set of the N classes. We learn the sub-set of classes from where to transfer, and how much to transfer from each of them, via an estimate of the Leave One Out (LOO) error on the training set. Determining how much to transfer helps avoiding negative transfer. Therefore, in case of non- informative prior knowledge, transfer might be disregarded completely (when to transfer).

We also investigated the possibility to develop an algorithm able to mimic the knowledge transfer across modalities that happens in biological systems. We made the working assumption that in a first stage the system have access to the audio-visual patterns on both modalities, and that the modalities are synchronous. Hence the system learns the mapping between each audio-visual couple of input data. The classifier is designed as follows: each modality is classified separately by a specific algorithm, and the outputs of these classifiers are then combined together to provide the final, multi-modal classification. When the classifier receives an input data where one of the two modalities is very noisy, or completely missing, the internal model generates a 'virtual input' that replaces the noisy/missing one, and incrementally updates its internal representation. Experiments on different multimodal settings show that our algorithm improves significantly its performance in the presence of missing data in one of the two modalities, therefore demonstrating the usefulness of transfer of knowledge across modalities. Moreover the framework is deeply rooted on the theory of online learning that gives theoretical guarantees on the optimality of the approach.

5.3. Knowledge transfer in rodents

In this section we focus on our neuro-behavioral research in rodents demonstrating that trans-modal category transfer can be used as a mechanism to respond meaningfully to unexpected stimuli in a given sensory modality. Specifically we investigated whether and how, knowledge acquired during learning about the relevance of stimulus features in one sensory modality (here, audition) can be transferred to novel, unexpected stimuli of another sensory modality (here, vision).

We trained rodents (gerbils, Meriones unguiculatus) to associate a slow and a fast presentation rate of auditory tone pips with the Go response and NoGo response, respectively, in an active avoidance paradigm (shuttle box). After a predefined performance criterion was reached for the discrimination task in the auditory modality, a second training phase was initiated in which the sensory modality of the stimuli was changed from auditory to visual. For one animal group (congruent group) the contingency of the two stimulus presentation rates with the Go/Nogo- responses stayed the same irrespective of the modality of stimulation, for a second group (incongruent group) it was reversed across modalities.

After the modality switch, the congruent groups showed a higher acquisition rate of the conditioned responses than the incongruent groups indicating a crossmodal transfer of the rate-response association. During training, the electrocorticogram was recorded from two multitelectrode arrays chronically implanted onto the epidural surface of primary auditory and the visual cortex.

Cortical activity patterns from the ongoing electrocorticogram associated with the Go- and the NoGo stimuli were determined in the spatial distribution of signal power using a multivariate pattern classification procedure. For animals of the congruent group showing correct discrimination already during the first visual training sessions, we suspect that these individuals transferred the rate-response association learned during auditory training to the visual training. In these animals activity-patterns observed in the ongoing electrocorticogram of the auditory and the visual cortex were associated both with the auditory and the visual Go- and the NoGo-stimuli. We suggest that activity in both auditory and visual cortex was instrumental for achieving the cross-modal transfer of learned associations.

6. Application scenario

The one goal of the DIRAC project was to establish the paradigm of incongruency as a novel method of information retrieval within a model hierarchy. Research was inspired not only by theoretical questions, but also by the urge to find physiological evidence for special forms of hierarchies and models. With the incongruency reasoning formulated, fostered and the principles tested with first experiments conducted by the partners, a growing need for experimental data to further investigate and evaluate the findings led to the aggregation of different databases within the project. Over time, databases were assembled and used by the partners to drive their development and do verify their findings, as dscribed in Section 6.1. Evaluation is described in Sections 6.2 and 6.3.

6.1. Datasets

The following sub-sections give descriptions of the database of audio-visual recordings, the database of multichannel in-ear and behind the ear head related and bin-aural room impulse responses, the database of frequency modulated sweeps for STRF estimation, the database of OOV and OOL recordings and the database of motion captured actor walking on a treadmill.

Database of audio-visual recordings

Within the DIRAC project, two application domains were defined for rare and incongruent event detection, namely the security and surveillance on the one hand, and in-home monitoring of elderly people on the other.

Based on these application domains, scenarios have been developed by the project partners to show the potential of the DIRAC theoretical framework and the techniques developed within the project, while attempting to address realistic and interesting situations. Each scenario in turn was recorded by different partners using professional audio and video recording hardware assembled into two recording platforms, the AWEAR II and the OHSU recording setup. During the project, the partners refined both the developed detectors and the scene descriptions. That – over time – formed the audio-visual database of the DIRAC project. Several hundred recordings were recorded in more than 50 recording sessions at different locations.

The raw recording data had to be pre-processed, including format changes, projections to correct lens distortion, synchronization between video and audio, and preview video with reduced resolution, prior to the application of the detectors and model by the partners. The data was categorized using a list of keywords and time stamps to help partners to identify and search for different human actions contained in each recording. For evaluation purposes, the audio and video data of the recordings were annotated on a frame by frame basis, giving the pixel position of different body parts of the actor, e.g. the pixel position of the head and the upper and lower body, or the speech/non-speech classification of the recorded audio signal as time positions.

Database of OOV and OOL recordings

Two data sets of audio recordings have been produced. The first one is a set of utterances containing Out-Of-Vocabulary (OOV) words and non-speech sounds, and the second one contains English In-Language (IL) spontaneous speech featuring intermittent switches to a foreign language (Out-Of-Language – OOL).

Database of multichannel in-ear and behind the ear head related and binaural room impulse responses

An eight-channel database of head-related impulse responses (HRIR) and binaural room impulse responses (BRIR) was generated within the DIRAC project. The impulse responses (IR) were measured with three-channel behind-the-ear (BTE) hearing aids and an in-ear microphone at both ears of a human head and torso simulator. The scenes' natural acoustic background was also recorded in each of the real world environments for all eight channels. Overall, the present database allows for a realistic construction of simulated sound fields for hearing instrument research and, consequently, for a realistic evaluation of hearing instrument algorithms. The database is available online in the DIRAC project website and described in more detail in (Kayser et al, 2009a).

Database of frequency modulated sweeps for STRF estimation

The spectro-temporal receptive field (STRF) is a common way to describe which features are encoded by auditory and visual neurons. STRFs are estimated by relating stimuli, visual or auditory, to the evoked response ensemble of the neuron. Once an STRF has been estimated it can be used to predict the linear part of the response of the neuron to new stimuli.

A new class of stimuli was generated within the DIRAC project consisting of frequency modulated (FM) sweeps. FM sweeps are an alternative to dynamic moving ripples (DMR) that are often used for STRF estimation. Both stimulus classes are designed such that they sample a large portion of the neuron's input space while having very low autocorrelations which makes them suitable for STRF estimation. However, as there is one main feature to divide the sweep class into two, the up and the down sweep, this characteristic can be used to investigate context dependence of the STRF. We showed that neurons as well as local field potentials (LFPs) are sensitive to different sweep orientation (Vangeneugden J. et al, 2011). Hence, one could estimate STRFs using a varying number of up and down sweeps to realize different context conditions. The database and example scripts are available online in the DIRAC project website.

It has been shown that there are shortcomings in the STRF and its estimates cannot account for arbitrary stimuli. The impact on the scientific community is a potentially better description of the STRF. Futhermore, it is possible to realize different context conditions which may yield new insights into encoding strategies on the level of single neurons. In terms of DIRAC it would give a direct comparison if a stimuli is standard or deviant

Database of motion captured actor walking on a treadmill

We investigated how temporal cortical neurons encode actions differing in direction, both forward versus backward as different facing directions. The stimuli in this study were locomotory actions, i.e. displacement of a human body, but whereby the translational component is removed, thus resembling an actor as if locomoting on a treadmill. This was done in order to more accurately pinpoint the neural code subserving walking direction (walking left- or rightward

either in a forward or backward fashion), for which actual physical displacements of the body are erroneous in determining the neural code.

Computationally, coding between forward and backward versions differs with respect to which aspects of the actions are relevant compared to coding for different facing directions. All possible combinations amount to 16 actions in total. All stimuli have been made available as *.avi such that playback is not limited to certain devices.

6.2. Integrated algorithms

For the evaluation process, both audio and visual detectors provided by different project partners were evaluated against the ground truth annotation of the different databases of the DIRAC project. The following subsections provide an overview of each detector.

Acoustic Object Detection – Speech/Non-Speech Discrimination

This detector aims to classify parts of the one-channel audio signal of a recorded scene as either speech or non-speech using a pre-trained model. The detector produces a binary label output every 500 milliseconds, indicating whether speech was detected or not.

The model used for this detection is based on amplitude modulation features coupled with a support vector machine classifier back-end. Features used for classification are modulation components of the signal extracted by computation of the amplitude modulation spectrogram. By construction, these features are largely invariant to spectral changes in the signal, thereby allowing for a separation of the modulation information from purely spectral information, which in turn is crucial when discriminating modulated sounds such as speech from stationary backgrounds (Bach, J.-H. Et al, 2010d). The SVM back-end allows a very robust classification since it offers good generalization performance (see also Section 3.2).

Acoustic Localization detector

The acoustic localization detector analyzes a 2-channel audio signal of a recorded scene. It aims to give directional information of every acoustic object it detects within every time frame of 80 milliseconds. The output for every time frame is a vector of yes/no information for all 61 non-overlapping segments between 0 and 180 degrees of arrival with respect to the stereo microphone basis. The detector is capable of detecting multiple acoustic objects within one time frame, but not capable to classify any localized acoustic object.

The detector uses a correlation-based feature front-end and a discriminative classification back-end to classify the location-dependent presence or absence of acoustic sources in a given time frame. The features are computed on the basis of the generalized cross correlation (GCC) function between two audio input signals. The GCC is an extension of the cross power spectral

density function, which is given by the Fourier transform of the cross correlation. Support Vector Machines are employed to classify the presence or absence of a source at each angle. This approach enables the simultaneous localization of more than one sound source in each time-frame (see also Section 3.2).

Tracker Tree

The tracker tree processes video information on a frame by frame basis by utilizing multiple specialized models organized in a hierarchical way (see Fig. 8). Each model operates on a frame by frame basis and gives confidence measures for the action it was designed for, i.e. “sitting” for a sitting person, “walking” for a walking or standing person, and “picking” for a person picking up something (Nater, F. Et al, 2009).

Conversation Detector

The multi-modal Conversation Detector uses the DIRAC principle of incongruency detection to discriminate between normal conversation and unusual conversational behavior, e.g. a person talking to himself. The Detector operates on output data from three detectors presented in this document: the speech/non-speech classifier, the audio localizer, the acoustic localization detector, and the tracker tree’s person detector. The output signals are combined into a DIRAC incongruence model instantiating a part-whole relationship (see also Section 2.2 for description of the part-whole relationship model). Within the Conversation Detector, three models on the general level (PT: person tracker, AL: audio localizer, SC: speech classifier), are combined to one conjoint model; one model on the specific level uses the fused data input of each model on the general level (see Fig. 9). The model on the specific level (Cf) utilizes a linear support vector machine and operates on the same (albeit fused) input data as the models on the general level. An incongruency is detected when the conjoint model accepts the input as conversation, whereas the specific model does not.

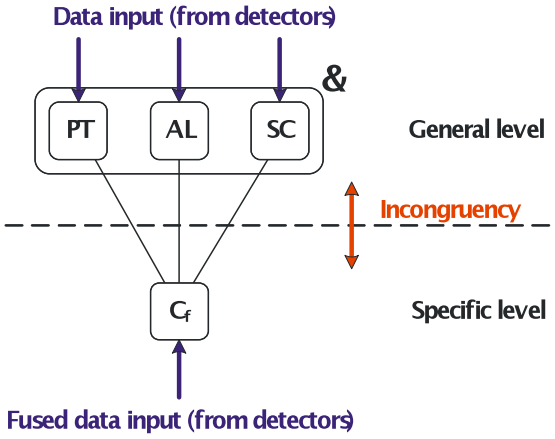

Figure 9: DIRAC incongruence model of the Conversation Detector.

Combined audio-visual incongruence detector

Combined audio-visual incongruence detector is an example of a conjunctive hierarchy in audio-visual processing. Alternative detectors (i.e. discriminative classifiers) were used to model events in a hierarchical manner, see Figure 10. We concentrate on the single audio-visual event of a human speaker in a scene and model it in two alternative ways. We assume a scene observed by a camera with wide view-field and two microphones. Visual processing detects the presence and position of a human. Sound processing detects the intensity of sound and its direction of arrival.

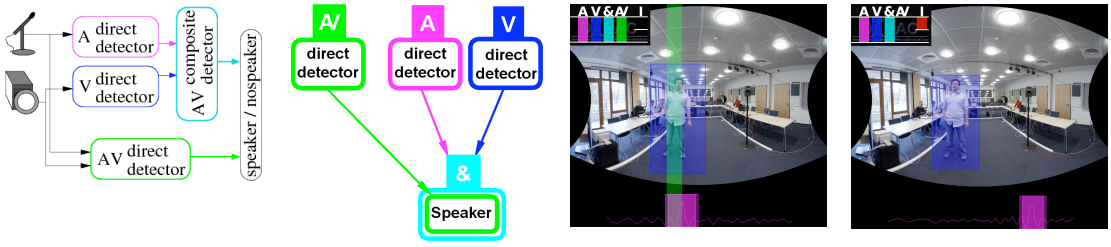

Figure10. Combined audio-visual incongruence detector. (a,b) Direct A and direct V detectors are combined into a composite A\V detector. Another direct A\V detector is run in parallel. Detectors are congruent (c) when they agree or incongruent (d) when the composite A\V detector is active while the direct A\V detector is passive.

The specific (direct) A\V is obtained by training a discriminative RBF SVM classifier on audio-visual features extracted from manually labelled training data of human speakers vs. background. It is evaluated on all spatial windows of meaningful size across the view-field, thus implicitly providing the positions of its decisions.

The general (composite) A\V detector is obtained by the conjunction of the direct visual detector and the direct audio detector. Unlike the direct A\V detector, it does not exploit the information about where the direct A and V detectors were active in the view-field. In effect, it looks whether they were active irrespectively of the position.

By construction, the composite A\V detector returns a positive outcome when observing a human body and human sound in different positions in the view-field. The direct A\V detector, on the other hand, is passive in this situation since it has been trained only on co-located human sound and visual examples. Thus, an incongruence appears. The appearance of such incongruence indicates a deficiency in the world model and can be used to initiate learning an updates of the model.

Combined tracker tree and transfer learning

We combined the tracker tree algorithm from for incongruent actions detection and the transfer learning in order to learn the new detected action. In this combination, the tracker tree detects an incongruent action, and asks for human annotation of few frames (from 1 to maximum 10) of it. These annotated frames are sent to the transfer learning algorithm, which learns the new action from the few annotated samples, exploiting the prior knowledge of the system. Note that the original tracker tree algorithm would need an average of 200 annotated samples for learning such action. Once the new class has been learned, the transfer learning method acts as an algorithmic annotators, and labels data sequences sent from the tracker tree where incongruent actions are detected. Once the number of annotation is of at least 200 frames, the data are sent back to the tracker tree, in order to build the new action representation and integrate it in the tree. The position where the action is added to the tree depends on where in the hierarchy the incongruency has been detected.

6.3. Results

The research and development cycle within the DIRAC project, starting with the development of ideas, the description and fostering of the DIRAC paradigm of incongruency and leading to the development of detectors and models as well as the aggregation of several databases is concluded with the evaluation of these models. The databases collected by the project partners have been prepared to serve as a basis for this evaluation by annotating its content.

Evaluation of audio-visual detectors

The following detectors have been evaluated against an evaluation database: the speech/non-speech detector, the acoustic localization detector, the “Walking”, “Sitting” and “Picking up” detectors from the tracker tree and the Conversation detector. The evaluation database was drawn from the pooled data of all project partners, sorted by content description and modality. All

recordings from the evaluation database were pre-processed with the DIRAC pre-processing pipeline and processed by the detector models. For the evaluation ground truth, each relevant scene has been annotated by human inspection.

One important fact should be mentioned explicitly: no recordings from the evaluation dataset were used to train the detector models. The partners only used their own recordings which have not been part of any of the databases collected within the DIRAC project. Thus, the selected evaluation dataset are completely new to the detector models.

The different detectors evaluated for both audio and video modality showed good performance over the evaluation data set from all three recording locations. The results per location never fell below 71% for a detector per location set, and there were only subtle differences in detector performance between recordings of different locations. Over all sets and all detectors, we correctly detected a remarkable average of 75.3% of all frames.

7. Summary and conclusions

The DIRAC project was motivated by the desire to bridge the gap between cognitive and engineering systems, and the observation that cognitive biological systems respond to unexpected events in a more robust and effective way. We first formulated a definition of rare events which goes beyond the traditional definition of event novelty. We then translated this conceptual framework to algorithms for a number of applications involving visual and auditory data. In parallel, we investigated learning mechanisms that can be used to learn the detected events. Finally, we developed a number of scenarios and collected data for which some of these algorithms could be tested with real data and in an integrated way.

8. Partner contribution

All partners participated in all the work described above, where different partners took the lead in different parts of the project. Thus partner CTU managed the work on visual sensors taking advantage of their expertise in fish-eye lenses, and the low-level representation of visual signals. Partner OL managed the work on auditory tasks and the low-level representation of auditory objects. Partner LIN lead the neurophysiological investigation into the nature of biological processing when confronting an unexpected event, using rodents. Partners ETHZ and KUL developed the system for the detection of incongruent events that is based on the detection of activities, where ETHZ was responsible for the compilation of the set of motion data and the development of the so-called tracker trees, while KUL managed the neurophysiological co-investigation using monkeys. The engineering group in KUL lead the design of the mobile AWEAR system. Partners OL and HUJI designed algorithms for the detection of rare novel

auditory and visual objects. Partners HUJI and IDIAP shared the work on developing a variety of learning algorithms that deal with rare novel events once they have been detected, developing methods for online learning and knowledge transfer. Partner BUT designed algorithms for the detection of OOV words. Partner FRA lead the design and management of the application scenario, while partner OHSU collected data in an independent living scenario. Both CTU and ETHZ have invested some resources in exploring applications other than independent living, especially during the final phase of the project.

9. References

9.1. Publications

Anemüller, J.; Bach, J.-H.; Gool, L. v.; Kayser, H.; Moreau, W. & Wabnik, S. (2010), Audio-visuelle Identifikation ungewöhnlicher Ereignisse - das DIRAC ProjektAudio-visual Identification of Unexpected Events – The DIRAC Project, *Fortschritte der Akustik – DAGA, 2010*.

Anemüller, J.; Schmidt, D.; Bach, J.-H. (2008), Detection of Speech Embedded in Real Acoustic Background Based on Amplitude Modulation Spectrogram Features. Proc. Interspeech, pp. 2582-2585.

Anemüller, J.; Bach, J.-H.; Caputo, B.; Havlena, M.; Jie, L.; Kayser, H.; Leibe, B.; Motlicek, P.; Pajdla, T.; Pavel, M.; Torii, A.; Gool, L. v.; Zweig, A. & Hermansky, H. (2008), The DIRAC AWEAR Audio-Visual Platform for Detection of Unexpected and Incongruent Events, *in* 'International Conference on Multimodal Interaction (ICMI)', pp. 289-293.

Anemüller, J.; Bach, J.-H.; Caputo, B.; Luo, J.; Ohl, F. W.; Orabona, F.; Vogels, R.; Weinshall, D. & Zweig, A. (2008), Biologically Motivated Audio-Visual Cue Integration for Object Categorization, *in* 'International Conference on Cognitive Systems (CogSys 2008).

Anemüller, J. (2007), Maximization of Component Disjointness: a Criterion for Blind Source Separation, *in* Mike E. Davies; Christopher J. James; Samer A. Abdallah & Mark D Plumbley, ed.,'Independent Component Analysis and Signal Separation', Springer, pp. 325--332.

Anemüller, J. (2006), Second-order Separation of Multidimensional Sources with Constrained Mixing System, *in* J. Rosca; Deniz Erdogmus; Jose C. Principe & S. Haykin, ed.,'Independent Component Analysis and Blind Signal Separation', Springer, , pp. 16--23.

Anemüller, J. (2006), 'Blind signal separation by disjoint component analysis', Technical report, University of Oldenburg.

AI, M.; Wetzel, W.; Scheich, H. & Ohl, F. W. (2007), Appetitive and Aversive Reinforcement and their Interaction During Auditory Learning, *in* 'Proc. 7-th Meeting German Neurosci. Soc.', pp. T31--1B.

Bach, J.-H. & Anemüller, J. (2010a), Detecting novel objects through classifier incongruence, *in* 'Interspeech', pp. 2206-2209.

Bach, J.-H.; Anemüller, J. & Kollmeier, B. (2010b), 'Robust Speech Detection in Real Acoustic Backgrounds with Perceptually Motivated Features', *Speech Comm*.

Bach, J.-H.; Kayser, H. & Anemüller, J. (2010c), Audio Classification and Localization for Incongruent Event Detection, *in* 'Workshop on the Detection and Identification of Rare Audiovisual Cues'.

Bach, J.-H.; Kollmeier, B. & Anemüller, J. (2010d), Modulation-based detection of speech in real background noise: Generalization to novel background classes, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 41-44.

Bach, J.-H. & Anemüller, J. (2009), 'Acoustic Object Detection in Adverse Conditions"NAG-DAGA 2009 International Conference on Acoustics', NAG-DAGA International Conference on Acoustics, Rotterdam, abstract and conference presentation, 315.

Bach, J.-H. & Anemueller, J. (2008), 'Hierarchical multi-class classification of sound data', *in* Jekosch & Hoffmann, ed.,'Fortschritte der Akustik - DAGA 2008', DEGA e.V. Berlin, 835-836.

Bach, J.-H. & Anemüller, J. (2008), 'Hierarchical approach to voice activity detection in realistic environmental noise', *The Journal of the Acoustical Society of America* **123**(5), 3331-3331.

Baene, W. D.; Premereur, E. & Vogels, R. (2007), 'Properties of shape tuning of macaque inferior temporal neurons examined using Rapid Serial Visual Presentation', *Journal of Neurophysiology* **97**, 2900--2916.

Bakstein, H. & Leonardis, A. (2007), Catadioptric Image-based Rendering for Mobile Robot Localization, *in* 'Proceedings of The Seventh Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras (OMNIVIS)', pp. 1--6.

Bakstein, H.; Havlena, M.; Pohl, P. & Pajdla, T. (2006), 'Omnidirectional sensors and their calibration for the Dirac project', Technical report, Czech Technical University, Center for Machine Perception.

Bar-Hillel, A. & Weinshall, D. (2007), Learning Distance Function by Coding Similarity, *in* 'Proceedings of the 24th international conference on Machine learning (ICML)', ACM", doi"http://doi.acm.org/10.1145/1273496.1273505, , pp. 65--72.

Bar-Hillel, A. & Weinshall, D. (2006), Subordinate class recognition using relational object models, *in* 'Advances in Neural Information Processing Systems (NIPS)', pp. 73--80.

Bar-Hillel, A. & Weinshall, D. (2006), 'Efficient Learning of Relational Object Class Models', *Tenth IEEE International Conference on Computer Vision (ICCV 2005)* **2**, 1762--1769.

Barber, D. (2007), 'Expectation Correction for smoothing in Switching Linear Gaussian State Space models', *Journal of Machine Learning Research* **7**, pp. 2515-2540.

Barber, D. (2006), 'A Novel Gaussian Sum Smoother for Approximate Inference in Switching Linear Dynamical Systems', *Journal of Machine Learning Research* **7**, 2515--2540.

Barber, D. (2006), 'Expectation Correction for Smoothed Inference in Switching Linear Dynamical Systems', *Journal of Machine Learning Research* **7**, 2515--2540.

Barber, D. & Chiappa, S. (2006), Unified Inference for Variational Bayesian Linear Gaussian State-Space Models, *in* '20th Conference on Neural Information Processing Systems, 4-7 Dec 2006, Vancouver, Canada'.

Barber, D. & Mesot, B. (2006), A Novel Gaussian Sum Smoother for Approximate Inference in Switching Linear Dynamical Systems, *in* Bernhard Schцlkopf; Thomas Hofmann & John Platt,

ed.,'Neural Information Processing Systems'.

Boenke, L.; Alais, D.; Ball, F. & Ohl, F. (2010), Inter-individual variability in audiovisual temporal-order judgments, *in* '11th International Multisensory Research Forum (IMRF)'.

Boenke, L.; Ball, F.; Bornfleth, M.; Deliano, M. & Ohl, F. (2010), The peak-latency of visual P1- and N1-components predicts perceived delays in an audiovisual temporal order judgment task, *in* '11th International Multisensory Research Forum (IMRF)'.

Boenke, L.; Ball, F.; Deliano, M. & Ohl, F. (2009), Different effects of within- and across-experiment variation of auditory and visual stimulus intensity on the conscious perception of temporal order, *in* '13th annual meeting of the Association for the Scientific Study of Consciousness (ASSC13)', pp. 59-60.

Boenke, L.; Deliano, M. & Ohl, F. (2009), Stimulus duration influences perceived simultaneity in audiovisual temporal order judgment, *in* '10th Annual Meeting of the International Multisensory Research Forum'.

Boenke, L.; Deliano, M. & Ohl, F. (2009), 'Stimulus duration influences perceived simultaneity in audiovisual temporal order judgment', *Experimental Brain Research* **198(2-3)**, 233-244.

Boenke, L. T.; Ohl, F. W.; Nikolaev, A. R.; Lachmann, T. & Van Leeuwen, C. (2009), 'Different time courses of Stroop and Garner effects in perception - An Event-Related Potentials Study', *NEUROIMAGE* **45(4)**(4), 1272-1288.

Boenke, L.; Deliano, M. & Ohl, F. (2008), Temporal aspects of auditory and visual stimuli processing assessed by temporal order judgment and reaction times, *in* 'International Multisensory Research Forum (IMRF)'.

Boenke, L.; Deliano, M. & Ohl, F. (2008), Neuronal correlates of spatial audio-visual temporal order perception, *in* 'International Multisensory Research Forum (IMRF)'.

Boenke, L. T.; Deliano, M. & Ohl, F. W. (2007), Interstimulus interaction in an audiovisual temporal order judgment task: The role of offset of the preceding stimulus and stimulus length on temporal perception, *in* '8th Annual Meeting of the International Multisensory Research Forum, July 5 - 7, 2007'.

Boenke, L. T.; Deliano, M. & Ohl, F. W. (2007), Role of stimulus length, intensity, and offset-onset interaction in audiovisual temporal order judgement, *in* '15th ESCoP Conference, Auditory perception'.

Budinger, E.; Scheich, H. & Ohl, F. (2008), 'Anatomical substrates for the processing of multimodal information via the primary auditory cortex"Frontiers in Human Neuroscience', 10th International Conference on Cognitive Neuroscience.

Budinger, E.; Jeschke, M.; Lenz, D. & Ohl, F. W. (2007), 'Gamma oscillations in gerbil auditory cortex during a target-discrimination task reflect matches with short-term memory', *Brain Research*.

Budinger, E.; Laszcz, A.; Lison, H.; Scheich, H. & Ohl, F. W. (2007), 'Non-sensory cortical and subcortical conncections of the primary auditory cortex in Mongolian gerbils. Bottom-up and top-

down processing of neuronal information via Field AI.', *Brain Research*.

Bujnak, M.; Kukelova, Z. & Pajdla, T. (2008), 'A general solution to the P4P problem for camera with unknown focal length"CVPR - Conference on Computer Vision and Pattern Recognition', CVPR - Conference on Computer Vision and Pattern Recognition.

Burget, L.; Brümmer, N.; Reynolds, D.; Kenny, P.; Pelecanos, J.; Vogt, R.; Castaldo, F.; Dehak, N.; Dehak, R.; Glembek, O.; Karam, Z.; Noecker, J. J.; Na, Y. H.; Costin, C. C.; Hubeika, V.; Kajarekar, S.; Scheffer, N. & ?ernocký, J. (2008), 'Robust Speaker Recognition Over Varying Channels', Johns Hopkins University, 81.

Burget, L.; Schwarz, P.; Matejka, P.; Hannemann, M.; Rastrow, A.; White, C.; Khudanpur, S.; Hermansky, H. & Cernocky, J. (2008), Combination of strongly and weakly constrained recognizers for reliable detection of OOVs, *in* 'International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', pp. 4.

Brümmer, N.; Strasheim, A.; Hubeika, V.; Matějka, P.; Burget, L. & Glembek, O. (2009), Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics, *in* 'Proc. Interspeech 2009', International Speech Communication Association, , pp. 2187--2190.

Byroed J., Kukelova Z., P. T. A. K. (2008), 'Fast and Robust Numerical Solutions to Minimal Problems for Cameras with Radial Distortion"CVPR - Conference on Computer Vision and Pattern Recognition', CVPR - Conference on Computer Vision and Pattern Recognition.

Castellini, C.; Tommasi, T.; Noceti, N.; Odone, F. & Caputo, B. (2011), 'Using object affordances to improve object recognition.', *IEEE Transaction on Autonomous Mental Development*.

Cernocky, J.; Szőke, I.; Hannemann, M. & Kombrink, S. (2010), Word-subword based keyword spotting with implications in OOV detection, *in* 'Asilomar Conference on Signals, Systems, and Computers', Institute of Electrical and Electronics Engineers, Pacific Grove, US, pp. 34.

Cernocky, J. (2009), Brno University of Technology detecting OOVs in DIRAC project, *in* 'SLTC newsletter 07/09 (Speech and Language Processing Technical Committee)IEEE Signal Processing Society'.

Cesa-Bianchi, N.; Gentile, C. & Orabona, F. (2009), Robust Bounds for Classification via Selective Sampling, *in* 'International Conference on Machine Learning (ICML)', ACM International Conference Proceeding Series, Montreal, pp. 121-128.

Chiappa, S. & Barber, D. (2007), Bayesian Linear Gaussian State Space Models for Biosignal Decomposition, *in* Bernhard Schцlkopf; John Platt & Thomas Hofmann, ed.,'Neural Information Processing Systems', MIT press, .

Chiappa, S. & Barber, D. (2007), 'Bayesian Linear Gaussian State Space Models for Biosignal Decomposition', *Signal Processing Letters* **14**, 267--270.

Cornelis, N.; Leibe, B.; Cornelis, K. & Gool, L. V. (2008), '3D Urban Scene Modeling Integrating Recognition and Reconstruction', *International Journal of Computer Vision* **78**, 121--141.

Cornelis, K.; Pajdla, T. & Havlena, M. (2007), Towards City Modeling from Omnidirectional

Video, *in* 'Proceedings of the 12th Computer Vision Winter Workshop (CVWW 2007)', pp. 123--130.

Cornelis, N.; Leibe, B.; Cornelis, K. & Gool, L. V. (2007), 3D City Modeling Integrating Recognition and Reconstruction, *in* '8th Conference on Optical 3D Measurement Techniques'.

Cornelis, N.; Cornelis, K. & Gool, L. V. (2006), Fast Compact City Modeling for Navigation Pre-Visualization, *in* 'Conference on Computer Vision and Pattern Recognition 2006', pp. 1339--1344.

Cornelis, N.; Leibe, B.; Cornelis, K. & Gool, L. V. (2006), 3D City Modeling using Cognitive Loops, *in* 'International Symposium on 3D Data Processing, Visualization and Transmission, 3DPVT 2006', pp. 9--16.

De Baene, W. & Vogels, R. (2010), 'Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials', *Cerebral Cortex* **20(9)**, 2145-65.

De Baene, W. & Vogels, R. (2009), 'Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials', *Cerebral Cortex*.

De Baene, W.; Ons, B.; Wagemans, J. & Vogels, R. (2008), 'Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons', *Learning and Memory* **15**, 717-727.

Deliano & Ohl (2009), 'Neurodynamics of category learning: Towards understanding the creation of meaning in the brain', *New Mathematics and Natural Computation (NMNC)* **5**, 61-81.

Deliano, M.; Scheich, H. & Ohl, F. (2009), 'Intracortical Microstimulation and its Role for Sensory Processing and Learning', *Journal of Neuroscience* **29(50)**, 15898-15909.

Deliano, M. & Ohl, F. W. (2007), States in the Ongoing Cortical Activity Carrying Information in Discrimination Learning of Differential Electrical Stimulation applied through a sensory cortex interface, *in* 'Proceedings of the 7th Meeting of the German Neuroscience Society/31st Gцttingen Neurobiology Conference, 2007', pp. T19--13B.

Deliano, M.; Ilango, A.; Fillbrandt, A.; Wetzel, W. & Ohl, F. W. (2006), Cortical activity associated with discrimination learning of auditory, audio-visual or intracortical electric stimuli after appe, *in* 'Int. Conf. Auditory Cortx. The Listening Brain, September 17-21 2006, Grantham, UK'.

Engelhorn, A.; Dann, B.; Deliano, M. & Ohl, F. W. (2007), Backward Masking Effects Produced by Intracortical Microstimulation in the Auditory Cortex, *in* 'Proc. 7-th Meeting German Neurosci. Soc.', pp. T19--14B.

Ess, A.; Leibe, B.; Schindler, K. & van Gool, L. (2008), A Mobile Vision System for Robust Multi-Person Tracking, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)'.

Ess, A.; Leibe, B. & Gool, L. V. (2007), Depth and Appearance for Mobile Scene Analysis, *in* 'International Conference on Computer Vision (ICCV'07)', pp. 1--8.

Feldman, D. & Weinshall, D. (2008), 'Motion Segmentation and Depth Ordering Using an

Occlusion Detector', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(7), 1171--1185.

Feldman, D. & Weinshall, D. (2006), Motion Segmentation Using an Occlusion Detector, *in* 'Workshop on Dynamical Vision, in 9th European Conference on Computer Vision (ECCV)', springer Berlin / Heidelberg, , pp. 34--47.

Fillbrandt, A. & Ohl, F. (2010), Patterns in the ongoing activity of the auditory and the visual cortex during audiovisual category transfer in rodents, *in* '11th International Multisensory Research Forum (IMRF)'.

Fillbrandt, A. & Ohl, F. (2009), Audiovisual category transfer in rodents, *in* '10th Annual Meeting of the International Multisensory Research Forum', pp. 34.

Fillbrandt, A. & Ohl, F. (2009), 'Cortical neurodynamics during audiovisual category transfer in rodents"3rd International Conference on Auditory Cortex', Conference on Auditory Cortex, Magdeburg, Germany, P 033.

Fillbrandt, A.; Zeghbib, A. & Ohl, F. (2009), 'Trial-to-trial variability of interaction dynamics between auditory and visual cortex during asynchronous audiovisual stimulation"8th Göttingen Meeting of the German Neuroscience Society', 8th Goettingen Meeting of the German Neuroscience Society, Goettingen, T23 - 3B (P. 1120).

Fillbrandt, A.; Deliano, M. & Ohl, F. (2008), 'Audiovisual category transfer in rodents, an electrophysiological study of directional influences between auditory and visual cortex"International Multisensory Research Forum (IMRF)', International Multisensory Research Forum (IMRF), 9th Annual Meeting.

Fillbrandt, A.; Deliano, M. & Ohl, F. (2008), 'Audiovisual category formation in rodents - an electrophysiological study on the dynamics of crossmodal interactions"Society for Neuroscience', Society for Neuroscience.

Fillbrandt, A.; Deliano, M. & Ohl, F. W. (2007), Audiovisual Category Transfer - an Electrophysiological Study in Rodents, *in* 'Proc. 7-th Meeting German Neurosci. Soc.', pp. T28--9B.

Fruend, I.; Naue, N.; Ohl, F. & Herrmann, C. (2008), Prestsimulus EEG activity in the alpha and beta/gamma bands is related to poststimulus performance, *in* 'Psychologie und Gehirn', pp. p. 119.

Fruend, I.; Ohl, F. & Herrmann, C. (2008), 'Spectral characteristics of mass activity in balanced inhibition networks"COSYNE - Computational and Systems Neuroscience', COSYNE - Computational and Systems Neuroscience.

Ganapathy, S.; Thomas, S. & Hermansky, H. (2009), 'Modulation frequency features for phoneme recognition in noisy speech', *Journal of the Acoustical Society of America* **125**(1), EL8-EL12.

Ganapathy, S.; Motlicek, P.; Hermansky, H. & Garudadri, H. (2008), Temporal Masking for Bit-rate Reduction in Audio Codec Based on Frequency Domain Linear Prediction, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)', pp. 4781--4784.

Ganapathy, S.; Motlicek, P.; Hermansky, H. & Garudadri, H. (2008), Spectral Noise Shaping: Improvements in Speech/Audio Codec Based on Linear Prediction in Spectral Domain, *in* 'Interspeech'.

Ganapathy, S.; Motlicek, P.; Hermansky, H. & Garudadri, H. (2008), 'Autoregressive Modeling of Hilbert Envelopes for Wide-Band Audio Coding"124th Convention of Audioengineering Society', 124th Convention of Audioengineering Society, Amsterdam.

Ganapathy, S.and Thomas, S. & Hermansky, H. (2008), Front-end for Far-ï¬eld Speech Recognition based on Frequency Domain Linear Prediction, *in* 'Interspeech'.

Gijsberts, A.; Tommasi, T.; Metta, G. & Caputo, B. (2010), Object recognition using visuo-affordance maps, *in* 'Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS 2010)'.

Goetze, S.; Rohdenburg, T.; Hohmann, V.; Kammeyer, K. D. & Kollmeier, B. (2007), Direction of Arrival Estimation based on the Dual Delay Line Approach for Binaural Hearing Aid Microphone Arrays, *in* 'Proc. Int. Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)', pp. 84--87.

Goldschmidt, J.; Wanger, T.; Engelhorn, A.; Friedrich, H.; Happel, M.; Ilango, A.; Engelmann, M.; Stuermer, I.; Ohl, F. & Scheich, H. (2009), 'High-resolution mapping of neuronal activity using the lipophilic thallium chelate complex TlDDC: Protocol and validation of the method', *Neuroimage* **49**, 303-315.

Goldschmidt, J.; Schildt, M.; Wetzel, W.; Zuschratter, W.; Laszcz, A.; Ohl, F. W.; Budinger, E.; Scheich, H. & Schulze, H. (2007), Structural Left-Right Asymmetries in Rodent Auditory Cortex, *in* 'Proc. 7-th Meeting German Neurosci. Soc.', pp. T19--1C.

Gool, L. V.; Jaeggli, T. & Koller-Meier, E. (2006), Combining Sample-Based and Analytical Density Propagation for Monocular Tracking, *in* 'IEEE CVPR workshop: Learning, Representation and Context for Human Sensing in Video', IEEE Computer Society, .

Hammer, R.; Brechmann, A.; Ohl, F.; Weinshall, D. & Hochstein, S. (2010), 'Differential category learning processes: The neural basis of learning by comparison', *Neuroimage*.

Hammer, R.; Brechmann, A.; Ohl, F.; Weinshall, D. & Hochstein, S. (2009), 'The neuronal basis of category learning by comparison"Neuroscience', Neuroscience, Annual Meeting PublicationsAbstr. 503.7.

Hammer, R.; Diesendruck, G.; Weinshall, D. & Hochstein, S. (2009), 'The development of category learning strategies: What makes the difference?"', *Cognition* **112(1)**, 105-119.

Hammer, R.; Hertz, T.; Hochstein, S. & Weinshall, D. (2009), 'Category Learning from Equivalence Constraints', *Cognitive Processing* **10(3)**, 211-232.

Hammer, R.; Bar-Hillel, A.; Hertz, T.; Weinshall, D. & Hochstein, S. (2008), 'Comparison Processes in Category Learning: From Theory to Behavior', *Brain Research* **1225**, 102-118.

Hammer, R.; Brechmann, A.; Ohl, F.; Diesendruck, G.; Weinshall, D. & Hochstein, S. (2008),

'Differential Learning Processes for Categorization"Vision Sciences Society', Vision Sciences Society 8th Annual Meeting, oral presentation.

Hammer, R.; Hertz, T.; Hochstein, S. & Weinshall, D. (2008), 'Category Learning from Equivalence Constraints', *Cognitive Processing*.

Hammer, R.; Diesendruck, G.; Weinshall, D. & Hochstein, S. (2007), The development of category learning strategies, *in* 'Perception (supplement), ECVP 2007', pp. 175.

Hammer, R.; Hertz, T.; Hochstein, S. & Weinshall, D. (2007), Classification with Positive and Negative Constraints: Theory, Computation and Cognition, *in* '2nd International Symposium on Brain, Vision and Artificial intelligence (BVAI)'.

Hammer, R.; Hertz, T.; Hochstein, S. & Weinshall, D. (2006), Category learning from positive and negative pairwise relations, *in* 'European Conference on Visual Perception'.

Hannemann, M. & et al (2010), Similarity scoring for recognized repeated Out-of-Vocabulary words, *in* 'Interspeech'.

Happel, M.; Jeschke, M. & Ohl, F. (2010), Afferent and intrinsic input systems defining spectral freqeuncy integration in primary auditory cortex, *in* '33rd Meeting of Assoc. Res. Otolaryngol'.

Happel, M.; Jeschke, M.; Handschuh, J.; Deliano, M. & Ohl, F. (2009), 'Parallel electrophysiological and behavioral analysis of layer-specific electrical microstimulation in primary auditory cortex - implications for the subcortical-loop hypothesis"3rd International Conference on Auditory Cortex', 3rd Int. Conf. on Auditory Cortex, Magdeburg, P044.

Happel, M.; Jeschke, M.; Handschuh, J.; Deliano, M. & Ohl, F. (2009), Parallel electrophysiological and behavioral analysis of layer-specific electrical microstimulation in primary auditory cortex - implications for the subcortical-loop hypothesis, *in* '8th Meeting of the German Neuroscience Society'.

Happel, M.; Jeschke, M.; Deliano, M. & Ohl, F. (2008), 'Comparison of laminar high resolution CSDs evoked by pure tones or ICMS stimulation in gerbil primary auditory cortex - implications for cortical neuroprostheses"Assoc. Res. Otolaryngol.', Assoc. Res. Otolaryngol., Abs.: 280, p. 96.

Happel, M.; Jeschke, M.; Handschuh, J.; Deliano, M. & Ohl, F. (2008), 'Parallel electrophysiological and behavioral analysis of layer-specific electrical microstimulation in primary auditory cortex - implications for the subcortical-loop hypothesis"Society for Neuroscience', Soc. Neurosci. Abstr.

Happel, M.; Mueller, S.; Anemueller, J. & Ohl, F. (2008), 'Predictability of STRFs in auditory cortex neurons depends on stimulus class"Interspeech', Interspeech 2008, ISCA, Brisbane.

Happel, M.; Jeschke, M.; Deliano, M. & Ohl, F. (2007), Spatial and Temporal Activity Characteristics in Primary Auditory Cortex Investigated with Current Source Density Analysis under Pharmacological Manipulation, *in* 'Proceedings of the 7th Meeting of the German Neuroscience Society'.

Havlena, M.; Heller, J.; Kayser, H.; Bach, J.-H.; Anemüller, J. & Pajdla, T. (2010), Incongruence

Detection in Audio-Visual Processing, *in* 'Workshop on the Detection and Identification of Rare Audiovisual Cues (ECML/PKDD)'.

Havlena, M.; Ess, A.; Moreau, W.; Torii, A.; Jancosek, M.; Pajdla, T. & Gool, L. v. (2009), AWEAR 2.0 System: Omni-directional Audio-Visual Data Acquisition and Processing, *in* '1st Workshop on Egocentric Vision, CVPR', pp. 49-56.

Havlena, M.; Torii, A.; Knopp, J. & Pajdla, T. (2009), Randomized Structure from Motion Based on Atomic 3D Models from Camera Triplets, *in* '2009 IEEE Conference on Computer Vision and Pattern Recognition'.

Havlena, M.; Torii, A. & Pajdla, T. (2009), Randomized Structure from Motion Based on Atomic 3D Models from Camera Triplets, *in* 'Computer Vision Winter Workshop'.

Havlena, M.; Heller, J.; Torii, A. & Pajdla, T. (2008), 'Omnidirectional Audio-Visual Data Acquisition and Processing', Technical report, CMP CTU Research Report, CMP-CTU-TR-2008-25.

Havlena, M.; Pajdla, T. s. & Cornelis, K. (2008), Structure from Omnidirectional Stereo Rig Motion for City Modeling, *in* 'VISAPP 2008 - Third International Conference on Computer Vision Theory and Applications', pp. 407-414.

Hendel, A.; Weinshall, D. & Peleg, S. (2010), Identifying Surprising Events in Videos Using Bayesian Topic Models, *in* 'Proceedings: 10th Asian Conference of Computer Vision (ACCV)'.

Hengel, P. v. & Anemüller, J. (2009), 'Audio Event Detection for In Home Care."NAG-DAGA 2009 International Conference on Acoustics', NAG-DAGA International Conference on Acoustics, Rotterdam, abstract conference presentation, 209.

Hengel, P. v.; Huisman, M. & Appell, J. (2009), *Sounds like trouble*, Shaker Publishing, chapter Human Factors - Security and Safety.

Hermansky, H. (2008), 'Dealing With Unexpected Words in Automatic Recognition of Speech', Technical report, Idiap Research Institut.

Hermansky, H.; Ganapathy, S.; Garudadri, H. & Motlicek, P. (2007), Non-uniform Speech/Audio Coding Exploiting Predictability of Temporal Evolution of Spectral Envelopes, *in* 'TSD', pp. 350--357.

Hermansky, H. & Motlicek, P. (2007), LP-TRAP based speech features for ASR, *in* '4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms'.

Hermansky, H. (2006), Detecting Information-Rich Events with Multiple Sensors, *in* 'CogSys II Conference', pp. 27.

Hermansky, H. (2006), A Brief Description of DIRAC project, *in* 'EuCognition Inaugural Meeting: The European Network for the Advancement of Artificial Cognitive Systems'.

Hermansky, H. (2006), Automatic Recognition of Speech Consistent with Some Properties of Auditory Cortical Receptive Fields, *in* .

Hermansky, H. (2006), Detecting Information-rich Events, *in* 'CogSys II, Radboud University, Nijmefen, NL, April 12-13 2006'.

Hermansky, H. (2006), Machine Recognition of Speech Consistent with some Properties of Auditory Cortical Receptive Fields, *in* 'International Conference on the Auditory Cortex'.

Herrmann, C.; Fruend, I. & Ohl, F. (2009), 'A biologically plausible network of spiking neurons can simulate human EEG responses"Bernstein Conference on Computational Neuroscience BCCN', Bernstein Conference on Computational Neuroscience BCCN, 160-161.

Herrmann, C. & Ohl, F.Sendhoff, B.; Koerner, E.; Sporns, O.; Ritter, H. & Doya, K., ed. (2009), *Cognitive adequacy in brain-like intelligence, Creating Brain-like Intelligence*, Vol. 5436/2009, Springer Verlag, Berlin, LNCS.

Hertz, T.; Bar-Hillel, A. & Weinshall, D. (2006), Learning a Kernel Function for Classification with Small Training Samples, *in* 'Proceedings of 23rd International Conference on Machine Learning (ICML)', pp. 401--408.

Hillel, A. B. & Weinshall, D. (2008), 'Efficient Learning of Relational Object Class Models', *International Journal of Computer Vision (IJCV)* **77**(1-3), 175--198.

Hollosi, D.; Schröder, J.; Goetze, S. & Appell, J.-E. (2010), Voice Activity Detection Driven Acoustic Event Classification for Monitoring in Smart Homes, *in* 'Submitted to 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies'.

Hollosi, D.; Wabnik, S.; Gerlach, S. & Kortlang, S. (2010), Catalog of Basic Scenes for Rare/Icongruent Event Detection, *in* .

Hurych, T.; Svoboda, J. & Trojanova, Y. (2009), Active Shape Model and Linear Predictors for Face Association Refinement, *in* '2009 IEEE 12th International Conference on Computer Vision Workshops'.

Ilango, A.; Wetzel, W.; Scheich, H. & Ohl, F. (2009), 'The combination of appetitive and aversive reinforcers and the nature of their interaction during auditory learning', *Neuroscience*.

Jaeggli, T.; Koller-Meier, E. & Gool, L. V. (2007), Learning Generative Models for Monocular Body Pose Estimation, *in* 'Asian Conference on Computer Vision (ACCV'07)'.

Jaeggli, T.; Koller-Meier, E. & Gool, L. V. (2007), Multi-Activity Tracking in LLE Body Pose Space, *in* '2nd Workshop on HUMAN MOTION Understanding, Modeling, Capture and Animation', Springer Berlin / Heidelberg, , pp. 42--57.

Jaeggli, T.; Koller-Meier, E. & Gool, L. V. (2006), Monocular Tracking with a Mixture of View-Dependent Learned Models, *in* '4th International Conference on Articulated Motion and Deformable', Springer, , pp. 494--503.

Jancosek, M. & Pajdla, T. (2008), Effective seed generation for 3D reconstruction, *in* '13th Computer Vision Winter Workshop'.

Jeschke, M.; Deliano, M.; Fillbrandt, A.; Freeman, W. & Ohl, F. (2008), 'Neurodynamcis in auditory cortex during learning"Congress of the International Organization of Psychophysiology',

Congress of the International Organization of Psychophysiology, St. Petersburg, Russia.

Jeschke, M. & F.W., O. (2008), 'Effect of training paradigms on behavioral strategy in an auditory discrimination task"Society for Neuroscience', Soc. Neurosci., Abstr. 850.16.

Jeschke, M.; Lenz, D.; Budinger, E.; Herrmann, C. & F.W., O. (2008), 'Gamma oscillations in gerbil auditory cortex during a target-discrimination task reflect matches with short-term memory', *Brain Res* **1220**, 70-80.

Jeschke, M.; Lenz, D.; Herrmann, C. & Ohl, F. (2008), 'Gamma oscillations during an auditory target-discrimination task reflect matches with short-term memory - A parallel study in humans and rodents"Assoc. Res. Otolaryngol', Assoc. Res. Otolaryngol, Abs.: 892, p. 304.

Jeschke, M.; Deliano, M. & Ohl, F. W. (2007), Disentangling the contribution of intracortical and thalamo-cortical projections to the generation of subthreshold spectral receptive fields in the auditory cortex, *in* 'Proc. 7-th Meeting German Neurosci. Soc.', pp. TS16--16C.

Jie, L.; Orabona, F.; Caputo, B. & Cesa-Bianchi, N. (2010), OM-2: an online multi-class multi kernel learning algorithm, *in* 'Proceedings of the fourth IEEE Online Learning for Computer Vision Workshop, in conjunction with CVPR 2010'.

Jie, L.; Orabona, F. & Caputo, B. (2009), 'An online framework for learning novel concepts over multiple cues', *Proceedings of Asian Conference on Computer Vision (ACCV)* **1**, 1-12.

Jimison, M. P.; Weinshall, D.; Zweig, A.; Ohl, F. W. & Hermansky, H. (2008), 'Detection and Identification of Rare Events in Cognitive and Engineering Systems', Technical report, Idiap.

Kaliukhovich, D. & Vogels, R. (2011), 'Stimulus Repetition Probability Does Not Affect Repetition Suppression in Macaque Inferior Temporal Cortex', *Cerebral Cortex* **in press**.

Karafiát, M. (2009), 'Study of linear transformations applied to training of cross-domain adapted large vocabulary continuous speech recognition systems'.

Kayser, H. (2010), 'Leblose Lebensretter (Lifeless life-saver)', radio interview, broadcasted on "Deutschlandfunk", programme: "Forschung aktuell" (Research today).

Kayser, H.; Ewert, S. D.; Anemüller, J.; Rohdenburg, T.; Hohmann, V. & Kollmeier, B. (2009a), 'Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses', *EURASIP Journal on Advances in Signal Processing* **2009**, 1-10.

Kayser, H.; Kollmeier, B. & Anemüller, J. (2009b), 'Blind and Non-Blind Spatial Signal Processing Using Head-Related Impulse Responses"NAG-DAGA International Conference on Acoustics', NAG/DAGA International Conference on Acoustics, Rotterdam, abstract conference presentation.

Keshet, J.; Grangier, D. & Bengio, S. (2007), Discriminative Keyword Spotting, *in* 'Workshop of Non-Linear Speech Processing (NOLISP)', Springer Berlin / Heidelberg, , pp. 220--229.

Ketabdar, H.; Hannemann, M. & Hermansky, H. (2007), Detection of Out-of-Vocabulary Words in Posterior Based ASR, *in* '8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)', pp. 1757--1760.

Ketabdar, H. & Hermansky, H. (2006), 'Identifying unexpected words using in-context and out-of-context phoneme posteriors'(68), Technical report, IDIAP.

Kliper, R.; Serre, T.; Weinshall, D. & Nelken, I. (2010), The Story of A Single Cell: Peeking into the Semantics of Spikes, *in* 'Proceedings: IAPR Workshop on Cognitive Information Processing (CIP)'.

Kliper, R.; Vaizman, Y.; Weinshall, D. & Portuguese, S. (2010), Evidence for depression and schizophrenia in speech prosody, *in* 'Proceedings: Second ISCA Tutorial and Research Workshop on Experimental Linguistics'.

Knopp, J.; Prasad, M.; Willems, G.; Timofte, R. & Gool, L. v. (2010), 'Hough Transform and 3D SURF for rbust three dimensional classification"11th European Conference on Computer Vision (ECCV)', 11th European Conference on Computer Vision (ECCV), accepted.

Knopp, J.; Sivic, J. & Pajdla, T. (2009), 'Location recognition using large vocabularies and fast spatial matching', Technical report, Willow CTU Research report.

Kombrink, S. (2010), OOV detection and beyond, *in* 'DIRAC workshop at ECML/PKDD'.

Kombrink, S.; Hanneman, M.; Burget, L. & Hermansky, H. (2010), Recovery of rare words in lecture speech, *in* 'accepted for TSD 2010 (13th International Conference on Text, Speech and Dialogue)', Lecture Notes in Articial Intelligence (LNAI), Springer-Verlag, Brno, Czech Republic.

Kombrink, S.; Burget, L.; Matejka, P.; Karafiat, M. & Hermansky, H. (2009), 'Posterior-based Out of Vocabulary Word Detection in Telephone Speech', *in* ISCA, ed.,'Interspeech 2009', Interspeech 2009, Brighton, GB, ISSN 1990-9772, 80-83.

Kukelova, Z.; Bujnak, M. & T., P. (2008), 'Polynomial eigenvalue solutions to the 5-pt and 6-pt relative pose problems"BMVC - British Machine Vision Conference', BMVC 2008.

Kukelova, Z.; Bujnak, M. & T., P. (2008), 'Automatic Generator of Minimal Problem Solvers"ECCV - European Conference on Computer Vision', ECCV.

Kukelova, Z. & Pajdla, T. (2007), A minimal solution to the autocalibration of radial distortion, *in* 'Proceedings of Computer Vision and Pattern Recognition Conference (CPVR)', IEEE Press, , pp. 1--7.

Kukelova, Z. & Pajdla, T. (2007), Two Minimal Problems for Cameras with Radial Distortion, *in* 'Proceedings of The Seventh Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras (OMNIVIS)', pp. 1--8.

Kukelova, Z. & Pajdla, T. (2007), Solving polynomial equations for minimal problems in computer vision, *in* 'Proceedings of the 12th Computer Vision Winter Workshop (CVWW 2007)', pp. 12--19.

Kurt, S.; Deutscher, A.; Crook, J. M.; Ohl, F. W.; Budinger, E.; Moeller, C. K.; Scheich, H. & Schulze, H. (2008), 'Auditory cortical contrast enhancing by global winner-take-all inhibitory interactions', *PLoS ONE*.

Kurt, S.; Crook, J. M.; Ohl, F. W.; Scheich, H. & Schulze, H. (2006), 'Differential effects of iontophoretic in vivo application of the GABAA antagonists bicuculline and gabazine in sensory cortex', *Hearing Research* **212**, 224--235.

Leibe, B.; Schindler, K.; Cornelis, N. & Gool, L. V. (2008), 'Coupled Object Detection and Tracking from Static Cameras and Moving Vehicles', **30(1)**, 1683--1698.

Leibe, B.; Cornelis, N.; Cornelis, K. & Gool, L. V. (2007), Dynamic 3D Scene Analysis from a Moving Vehicle, *in* , pp. 1--8.

Leibe, B.; Schindler, K. & Gool, L. V. (2007), Coupled Detection and Trajectory Estimation for Multi-Object Tracking, *in* 'International Conference on Computer Vision (ICCV'07)', pp. 1--8.

Leibe, B.; Cornelis, N.; Cornelis, K. & Gool, L. V. (2006), Integrating Recognition and Reconstruction for Cognitive Traffic Scene Analysis from a Moving Vehicle, *in* 'DAGM Annual Pattern Recognition Symposium', Springer Berlin / Heidelberg, , pp. 192--201.

Leibe, B.; Mikolajczyk, K. & Schiele, B. (2006), Segmentation Based Multi-Cue Integration for Object Detection, *in* 'British Machine Vision Conference (BMVC'06)', pp. III:1169.

Lenz, D.; Jeschke, M.; Schadow, J.; Naue, N.; Ohl, F. W. & Herrmann, C. (2007), 'Human EEG very high frequency oscillations reflect the number of matches with a template in auditory short-term memory', *Brain Research* **Epub**.

Lippert, M.; Steuddel, T.; Ohl, F.; Logothetis, N. & Kayser, C. (2010), 'Coupling of neural activity and fMRI-BOLD in the motion area MT', *Magnetic Resonance Imaging*.

Lovitt, A.; Pinto, J. P. & Hermansky, H. (2007), 'On Confusions in a Phoneme Recognizer', Technical report, IDIAP.

Luo, J.; Caputo, B.; Zweig, A.; Bach, J.-H. & Anemueller, J. (2008), Object Category Detection using Audio-visual Cues, *in* 'International Conference on Computer Vision Systems (ICVS08)'.

Luo, J.; Pronobis, A. & Caputo, B. (2007), SVM-Based Transfer of Visual Knowledge Across Robotic Platforms, *in* '5th International Conference on Computer Vision Systems (ICVS07)'.

Luo, J.; Pronobis, A.; Caputo, B. & Jensfelt, P. (2007), Incremental Learning for Place Recognition in Dynamic Environments, *in* 'IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 07)', pp. 721--728.

Maganti, H. K.; Gatica-Perez, D. & McCowan, I. (2007), 'Speech Enhancement and Recognition in Meetings with an Audio-Visual Sensor Array', *IEEE Transactions on Audio, Speech and Language Processing* **15(8)**, 2257 - 2269.

Maganti, H. K.; Motlicek, P. & Gatica-Perez, D. (2007), Unsupervised Speech/Non-speech Detection for Automatic Speech Recognition in Meeting Rooms, *in* 'IEEE international conference on acoustics, speech, and signal processing (icassp)', pp. IV-1037--IV-1040.

Maganti, H. K. & Gatica-Perez, D. (2006), Speaker Localization for Microphone Array-Based ASR: The Effects of Accuracy on Overlapping Speech, *in* 'Proceedings of International

Conference on Multimodal Interfaces (ICMIâ™06)', pp. 35--38.

Martinec, D. & Pajdla, T. (2007), Robust Rotation and Translation Estimation in Multiview Reconstruction, *in* 'Proceedings of the Computer Vision and Pattern Recognition conference (CVPR)', IEEE Computer Society, , pp. 1--8.

Martinec, D. & Pajdla, T. (2006), 3D Reconstruction by Gluing Pair-wise Euclidean Reconstructions, or ``How to Achieve a Good Reconstruction from Bad Images'', *in* M. Pollefeys & K. Daniilidis, ed.,'Third International Symposium on 3D Data Processing, Visualization and Transmission 2006', IEEE Computer Society Press, , pp. 25--32.

Martinez-Gomez, J. & Caputo, B. (2011), 'Towards Semi-Supervised Learning of Semantic Spatial Concepts', *Journal of Physical Agents*.

Martinez-Gomez, J. & Caputo, B. (2011), Towards Semi-Supervised Learning of Semantic Spatial Concepts, *in* 'Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2011)'.

Mesot, B. & Barber, D. (2007), A Bayesian Treatment of Gain Adaptation in Switching AR-HMMs, *in* 'ICASSP 2007'.

Meyer, A.; Happel, M.; Ohl, F. & Anemüller, J. (2009), 'Estimation of spectro-temporal receptive fields based on linear support vector machine classification"3rd International Conference on Auditory Cortex', 3rd Int. Conf. on Auditory Cortex, Magdeburg, P068.

Meyer, A.; Happel, M.; Ohl, F. & Anemüller, J. (2009), 'Estimation of spectro-temporal receptive fields based on linear support vector machine classification', *BMC Neuroscience* **10**(10(Suppl 1)), 147.

Micusik, B. & Pajdla, T. (2006), 'Structure from Motion with Wide Circular Field of View Cameras', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(7), pp. 1135--1149.

Mikolov, T.; Karafiát, M.; Burget, L.; Černocký, J. & Khudanpur, S. (2010), Recurrent neural network based language model, *in* 'Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)', International Speech Communication Association, , pp. 1045--1048.

Motlicek, P.; Ganapathy, S.; Hermansky, H. & Garudadri, H. (2008), Temporal Masking for Bit-rate Reduction in Audio Codec Based on Frequency Domain Linear Prediction, *in* 'IEEE Int. Conf. on Acoustics, Speech, and Signal Processing'.

Motlicek, P. (2007), 'LP-TRAPs in all senses'(66), Technical report, IDIAP.

Motlicek, P.; Ganapathy, S.; Hermansky, H. & Garudadri, H. (2007), 'Non-uniform QMF Decomposition for Wide-band Audio Coding based on Frequency Domain Linear Prediction'(43), Technical report, IDIAP.

Motlicek, P.; Ullal, V. & Hermansky, H. (2007), Wide-Band Perceptual Audio Coding based on Frequency-domain Linear Prediction, *in* 'Proc. of ICASSP 2007', pp. I-265--I-268.

Motlicek, P.; Harinath, G.; Srinivasamurthy, N. & Hermansky, H. (2006), Speech Coding based on Spectral Dynamics, *in* 'Ninth International Conference on TEXT, SPEECH and DIALOGUE (TSD)', Springer Berlin / Heidelberg, , pp. 471--478.

Motlicek, P.; Hermansky, H.; Garudadri, H. & Srinivasamurthy, N. (2006), 'Narrow-Band audio Coding Exploiting Predictability of Temporal Evolution of Spectral Envelopes'(6-30), Technical report, IDIAP.

MFK, H.; Jeschke, M.; Deliano, M. & Ohl, F. W. (2007), Spatial and temporal activity characteristics in primary auditory cortex investigated with current source density analysis under pharmacological manipulation, *in* 'Proc. 7-th Meeting German Neurosci. Soc.', pp. TS16--15C.

Nater, F.; Grabner, H. & Gool, L. V. (2010), Visual abnormal event detection for prolonged independent living, *in* 'mHealth Workshop at IEEE Healthcom'.

Nater, F.; Grabner, H. & Gool, L. V. (2010), Exploiting Simple Hierarchies for Unsupervised Human Behavior analysis, *in* 'IEEE Computer Society Conference on Computer Vision and Pattern Recognition'.

Nater, F.; Grabner, H.; Jaeggli, T. & Van Gool, L. (2010), 'Tracker trees: hierarchies to spot rare events', 4th International Conference on Cognitive Systems (CogSys 2010) - Poster.

Nater, F.; Vangeneugden, J.; Grabner, H.; Gool, L. v. & Vogels, R. (2010), Discrimination of locomotion direction at different speeds: A comparison between macaque monkeys and algorithms, *in* 'ECML Workshop on rare audio-visual cues'.

Nater, F.; Grabner, H.; Jaeggli, T. & Gool, L. v. (2009), Tracker trees for unusual event detection, *in* 'ICCV 2009 Workshop on Visual Surveillance'.

Noblejas, M.; Wetzel, W.; Schulz, A. & Ohl, F. (2008), 'Differential effects of lesions of the anterior cingulate cortex or lesions of the orbitofrontal cortex on extinction, spontaneous recovery and reinstatement of an avoidance response"Society for Neuroscience', Soc. Neuosci., Abstr. 487.19.

Noceti, N.; Caputo, B.; Castellini, C.; Baldassarre, L.; Barla, A.; Rosasco, L.; Odone, F. & Sandini, G. (2009), Towards a theoretical framework for learning multi-modal patterns for embodied agents, *in* 'LNCS - Image Analysis and Processing (ICIAP 2009)', Springer Berlin / Heidelberg, , pp. 239-248.

Ohl, F. (2009), Neurodynamics in auditory cortex during categorization and rare-event processing, *in* 'Neuromorphic Workshop'.

Ohl, F. & Scheich, H.Holscher, C. & Munk, M., ed. (2009), *The role of neuronal populations in auditory cortex for learning - Information processing by neuronal populations*, Cambridge University Press.

Ohl, F.; Deliano, M.; Fillbrandt, A.; Freeman, W. & Scheich, H. (2008), 'Neurodynamics in auditory cortex during category learning"Frontiers in Human Neuroscience', 10th International Conference on Cognitive Neuroscience.

Ohl, F. W. & Scheich, H. (2007), 'Chips in your head', *Scientific American Mind* **April**, 64--69.

Ohl, F. W. & Scheich, H. (2006), 'Neuroprothetik. Hitech im Gehirn', *Gehirn und Geist* **10**, 64--67.

Orabona, F.; Jie, L. & Caputo, B. (2010), Online-Batch Strongly Convex Multi Kernel Learning, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR10)'.

Orabona, F.; Caputo, B.; Fillbrandt, A. & Ohl, F. (2009), A Theoretical Framework for Transfer of Knowledge Across Modalities in Artificial and Biological Systems, *in* 'IEEE 8th International Conference on Development and Learning, 2009. ICDL 2009.'.

Orabona, F.; Castellini, C.; Caputo, B.; Fiorilla, A.; E. & G., S. (2009), Model adaptation with least-squares SVM for adaptive hand prosthetics, *in* 'International Conference on Robotics and Automation (ICRA) 2009'.

Orabona, F.; Castellini, C.; Caputo, B.; Luo, J. & Sandini, G. (2009), 'Towards Life-long Learning for Cognitive Systems: Online Independent Support Vector Machine', *Pattern Recognition* **43**(4), 1402-1412.

Orabona, F.; Keshet, J. & Caputo, B. (2009), 'Bounded kernel-based perceptrons', *Journal of Machine Learning Research* **10**, 2643в€’2666.

Orabona, F.; Keshet, J. & Caputo, B. (2008), The projectron: a bounded kernel-based perceptron, *in* '25th International Conference on Machine Learning'.

Orabona, F.; Castellini, C.; Caputo, B.; Luo, J. & Sandini, G. (2007), Indoor Place Recognition using Online Independent Support Vector Machines, *in* 'Proceedings of the 18th British Machine Vision Conference (BMVC)', pp. 1090--1099.

Pajdla, T.; Havlena, M. & Heller, J. (2010), Learning from Incongruences, *in* 'DIRAC Workshop at ECML/PKDD 2010'.

Pajdla, T.; Havlena, M.; Heller, J.; Kayser, H.; Bach, Jц.-H. & Anemüller, Jц. (2009), 'Incongruence Detection for Detecting, Removing, and Repairing Incorrect Functionality in Low-Level Processing'(CTU--CMP--2009--19), Technical report, CTU Research Report.

Pajdla, T.; Heller, J.; Torii, A. & Havlena, M. (2008), 'Incongruence detection on sensor and low processing level', Technical report, CTU Research Report, CMP-CTU-TR-2008-27, Dec.

Pajdla, T. & Micusik, B. (2007), Multi-label image segmentation via max-sum solver, *in* 'Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)', IEEE Press, , pp. 1--6.

Parthasarathi, S. H. K.; Motlicek, P. & Hermansky, H. (2008), 'Exploiting contextual information for speech/non-speech detection"LNCS - Text, Speech and Dialogue', 451-459.

Pinto, J. & Hermansky, H. (2008), Combining Evidence from a Generative and a Discriminative Model in Phoneme Recognition, *in* 'Interspeech'.

Pinto, J. P.; Bayya, Y.; Hermansky, H. & Magimai-Doss, M. (2008), Exploiting Contextual

Information for Improved Phoneme Recognition, *in* 'IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)', pp. 4449--4452.

Pinto, J. P.; Bourlard, H.; Greve, Z. D. & Hermansky, H. (2007), 'Comparing Different Word Lattice Rescoring Approaches Towards Keyword Spotting'(32), Technical report, IDIAP.

Pinto, J. P.; Lovitt, A. & Hermansky, H. (2007), Exploiting Phoneme Similarities in Hybrid HMM-ANN Keyword Spotting, *in* 'Interspeech 2007, 8th Annual Conference of the International Speech Communication Association'.

Prasanna, S. R. M. & Hermansky, H. (2007), MRASTA and PLP in Automatic Speech Recognition, *in* '8th Annual Conference of the International Speech (INTERSPEECH 2007), August 27-31, 2007', pp. 1166--1169.

Prasanna, S. R. M.; Yegnanarayana, B.; Pinto, J. P. & Hermansky, H. (2007), 'Analysis of Confusion Matrix to Combine Evidence for Phoneme Recognition'(27), Technical report, IDIAP.

Prasanna, S. R. M. & Hermansky, H. (2006), 'Multi-RASTA and PLP in Automatic Speech Recognition', Technical report, Idiap.

Pronobis, A.; Martinez Mozos, O. & Caputo, B. (2008), SVM-based discriminative accumulation scheme for place recognition, *in* 'IEEE International Conference on Robotics and Automation'.

Quinn, M.; Jimison, H.; Weinshall, D.; Ohl, F.; Hermansky, H. & Pavel, M. (2010), Detection of Rare Events and Categorization, *in* '43rd Annual Meeting of the Society for Mathematical Psychology'.

Rohdenburg, T.; Goetze, S.; Hohmann, V.; Kammeyer, K.-D. & Kollmeier, B. (2008), Objective Perceptual Quality Assessment for Self-Steering Binaural Hearing Aid Microphone Arrays, *in* 'IEEE Int. Conf. on Acoustics, Speech and Signal Processing'.

Rohdenburg, T.; Hohmann, V. & Kollmeier, B. (2007), Robustness analysis for multi-channel hearing aid algorithms with binaural output by means of objective perceptual quality measures, *in* 'DAGA', DEGA, , pp. 365--366.

Rohdenburg, T.; Hohmann, V. & Kollmeier, B. (2007), Robustheitsanalyse von mehrkanaligen binauralen Hoergeraetealgorithmen mit Hilfe von psychoakustischen Bewertungsmassen, *in* 'Tagung der Deutschen Arbeitsgemeinschaft fuer Akustik (DAGA)'.

Rohdenburg, T.; Hohmann, V. & Kollmeier, B. (2007), Robustness analysis of binaural hearing aid beamformer algorithms by means of objective perceptual quality measures, *in* 'IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)', pp. 315--318.

Rohdenburg, T.; Hohmann, V. & Kollmeier, B. (2006), Subband-based Parameter Optimization in Noise Reduction Schemes by means of Objective Perceptual Quality Measures, *in* 'International Workshop on Acoustic Echo and Noise Control (IWAENC 2006)'.

Schaer, T.; Ewert, S.; Anemueller, J. & Kollmeier, B. (2009), 'Measurement, modelling and compensation of nonlinearities in hearing aid receivers"NAG-DAGA International Conference on Acoustics 2009', NAG/DAGA International Conference on Acoustics, Rotterdam, abstract

conference presentation (P. 208-209).

Scheich, H.; Brechmann, A.; Brosch, M.; Budinger, E. & Ohl, F. W. (2007), 'The cognitive auditory cortex: Task-specificity of stimulus representations', *Hearing Research* **229**(1-2), 213--224.

Schmidt, D. & Anemueller, J. (2007), Acoustic Feature Selection for Speech Detection Based on Amplitude Modulation Spectrograms, *in* 'Fortschritte der Akustik: DAGA 2007', Deutsche Gesellschaft für Akustik (DEGA), , pp. 347--348.

Schröder, J.; Wabnik, S.; van Hengel, P. & Goetze, S. (2011), Detection and Classification of Acoustic Events for In-Home Care'To be published in Springer Lecture Notes in Computer Science (LNCS)', Springer, .

Schwarz, P. (2009), 'Phoneme recognition based on long temporal context'.

Shalit, U.; Weinshall, D. & Chechik, G. (2010), Online Learning in the Manifold of Low-Rank Matrices, *in* 'Proceedings: Advances in Neural Information Processing Systems (NIPS)'.

Shumake, J.; Ilango, A.; Scheich, H.; Wetzel, W. & Ohl, F. (2009), 'Acquisition and Retrieval of Avoidance Learning by the Lateral Habenula and Ventral Tegmental Area', *Journal of Neuroscience*.

Shumake, J.; Ilango, A.; Scheich, H.; Wetzel, W. & Ohl, F. (2009), Differential neuromodulation of acquisition and retrieval of avoidance learning by the lateral habenula and ventral tegmental area, *in* 'Society for Neuroscience'.

Sivaram, G. & Hermansky, H. (2008), 'Emulating Temporal Receptive Fields of Auditory Mid-brain Neurons for Automatic Speech Recognition"16th European Signal Processing Conference', 16th European Signal Processing Conference, Lausanne, Switzerland.

Sivaram, G. & Hermansky, H. (2008), Introducing Temporal Asymmetries in Feature Extraction for Automatic Speech Recognition, *in* 'Interspeech'.

Sivaram, G. S. & Hermansky, H. (2008), Emulating Temporal Receptive Fields of Higher Level Auditory Neurons for ASR, *in* '11th International Conference on Text, Speech and Dialogue'.

Sorkin, A.; Weinshall, D. & Peled, A. (2008), The distortion of reality perception in schizophrenia patients, as measured in Virtual Reality, *in* '16th Annual Medicine Meets Virtual Reality Conference (MMVR)'.

Sorkin, A.; Peled, A. & Weinshall, D. (2006), Detection of Inconsistent Audio-Visual Events in Virtual Reality, *in* 'European Conference on Visual Perception (ECVP)'.

Szöke, I.; Fapso, M.; Burget, L. & Cernocky, J. (2008), Hybrid Word-Subword Decoding for Spoken Term Detection, *in* 'SSCS 2008 - Speech search workshop at SIGIR', pp. 4.

Takagaki, K. & Ohl, F. (2009), Cortical plasticity of audiovisual mass action, *in* 'International Multisensory Research Forum'.

Takagaki, K.; Wanger, T. & Ohl, F. (2009), State-dependent patterns of interareal and intraareal

oscillatory coupling in gerbil auditory cortex, *in* '3rd International Conference on Auditory Cortex', pp. 144.

Takagaki, K.; Lippert, M.; Dann, B.; Wanger, T. & Ohl, F. (2008), 'Normalization of voltage-sensitive dye signal with functional activity measures', .

Takagaki, K. & Ohl, F. (2008), 'Cortical plasticity of audiovisual mass action"International Multisensory Research Forum', IMRF, 9th Annual Meeting, p. 98.

Takagaki, K.; Zhang, C.; Wu, J.-Y. & Lippert, M. T. (2008), 'Crossmodal propagation of sensory-evoked and spontaneous activity in the rat neocortex', *Neuroscience letters* **Volume 431, Issu**, 191--196.

Thomas, S.; Ganapathy, S. & Hermansky, H. (2008), 'Recognition Of Reverberant Speech Using Frequency Domain Linear Prediction', *IEEE Signal Processing Letters* **15**, 681-684.

Thomas, S.; Ganapathy, S. & Hermansky, H. (2008), Spectro-Temporal Features for Automatic Speech Recognition using Linear Prediction in Spectral Domain, *in* 'EUSIPCO'.

Thomas, S.; Ganapathy, S. & Hermansky, H. (2008), Hilbert Envelope Based Features for Far-Field Speech Recognition, *in* 'MLMI - Mashine Learning in Medical Imaging'.

Thomas, S.; Ganapathy, S. & Hermansky, H. (2008), Hilbert Envelope Based Spectro-Temporal Features for Phoneme Recognition in Telephone Speech, *in* 'Interspeech'.

Tommasi, T. & Caputo, B. (2010), Towards a quantitative definition of rareness, *in* 'Proc ECML2010 workshop on detection of rare audio visual events'.

Tommasi, T.; Orabona, F. & Caputo, B. (2010), Safety in numbers: learning categories from few examples with multi model knowledge transfer, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR10).'.

Tommasi, T. & Caputo, B. (2009), The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories, *in* 'British Machine Vision Conference (BMVC) 2009'.

Torii, A.; Havlena, M. & Pajdla, T. (2009), 'Camera Tracking and Autocalibration for Detecting and Correcting Camera De-Calibration'(CTU--CMP--2009--20), Technical report, CTU Research Report.

Torii, A.; Havlena, M. & Pajdla, T. (2009), 'Omnidirectional image stabilization by computing camera trajectory', *LNCS Computer Science (PSIVT'09) - Advances in Image and Video Technology* **5414**, 71-82.

Torii, A.; Havlena, M. & Pajdla, T. (2009), Omnidirectional Image Stabilization by Computing Camera Trajectory, *in* 'The 3rd Pacific-Rim Symposium on Image and Video Technology', pp. 71-82.

Torii, A.; Havlena, M. & Pajdla, T. (2009), 'From Google Street View to 3D City Models', *The ICCV 2009 IEEE Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras. Kyoto*.

Torii, A.; Havlena, M.; Jancosek, M.; Kukelova, Z. & Pajdla, T. (2008), 'Dynamic 3D Scene Analysis from Omni-Directional Video Data', Technical report, CMP CTU Research Report, CMP-CTU-TR-2008-26, Dec..

Torii, A.; Havlena, M.; Pajdla, T. & Leibe, B. (2008), Measuring camera translation by the dominant apical angle, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)'.

Torii, A. & Pajdla, T. (2008), Omnidirectional Camera Motion Estimation, *in* 'VISAPP - International Conference on Computer Vision Theory and Applications', pp. 577--584.

Torii, A. & Pajdla, T. (2006), 'Stereo Matching Insensitive to the Change of Camera Orientation', Technical report, Czech Technical University (CTU).

Ullah, M. M.; Orabona, F. & Caputo, B. (2009), You Live, You Learn, You Forget: Continuous Learning of Visual Places with a Forgetting Mechanism, *in* 'International Conference on Intelligent Robots and Systems 2009'.

Ullah, M. M.; Pronobis, A.; Caputo, B.; Luo, J.; Jensfelt, P. & Christensen, H. I. (2008), Towards Robust Place Recognition for Robot Localization, *in* 'IEEE International Conference on Robotics and Automation (ICRA08)'.

Ullal, V. & Motlicek, P. (2006), 'Audio coding based on long temporal segments: experiments with quantization of excitation signal'(46), Technical report, IDIAP.

Valente, F. & Hermansky, H. (2008), Hierarchical and Parallel Processing of Modulation Spectrum for ASR applications, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008),', pp. 4165--4168.

Valente, F. & Hermansky, H. (2008), On the Combination of Auditory and Modulation Frequency Channels for ASR applications, *in* 'Interspeech'.

Valente, F. & Hermansky, H. (2007), Combination of Acoustic Classifiers based on Dempster-Shafer Theory of evidence, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)'.

Valente, F.; Vepa, J.; Gollan, C.; Hermansky, H.; Schluter, R. & Plahl, C. (2007), Hierarchical Neural Networks Feature Extraction for LVCSR system, *in* 'Interspeech 2007, 8th Annual Conference of the International Speech Communication Association', pp. 42--45.

Valente, F.; Vepa, J. & Hermansky, H. (2007), Multi-stream Features Combination based on Dempster-Shafer Rule for LVCSR System, *in* .

Valente, F. & Hermansky, H. (2006), Discriminant Linear Processing of Time-Frequency Plane, *in* '9th International conference on Spoken Language Processing (INTERSPEECH 2006 - ICSLP)', Interspeech, , pp. 349--352.

Vangeneugden, J.; De Mazière, P.; Van Hulle, M.; Jaeggli, T.; Van Gool, L. & Vogels, R. (2011), 'Distinct Mechanisms for Coding of Visual Actions in Macaque Temporal Cortex', *Journal of Neuroscience* **31(2)**, 385-401.

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..

Vangeneugden, J.; Pollick, F. & Vogels, R. (2009), 'Functional differentiation of macaque visual temporal cortical neurons using a parametric action space', *Cerebral Cortex* **19(3)**, 593-611.

Vangeneugden, J.; Vancleef, K.; Jaeggli, T.; Gool, L. v. & Vogels, R. (2009), Coding of walking direction by macaque visual temporal cortical neurons, *in* 'Society for Neuroscience', Society for Neuroscience, .

Vangeneugden, J.; Vancleef, K.; Jaeggli, T.; Van Gool, L. & Vogels, R. (2009), 'Discrimination of locomotion direction in impoverished displays of walkers by macaque monkeys', *Journal of Vision*.

Vangeneugden, J.; Pollick, F. & Vogels, R. (2006), Responses of Macaque Superior Temporal Sulcus Neurons to a Parameterized Set of Dynamic Images of Visual Actions, *in* 'Society for Neuroscience 36th Annual meeting, Atlanta', pp. 547.1.

Verhoef, B. E.; Kayaert, G.; Franko, E.; Vangeneugden, J. & Vogels, R. (2008), 'Stimulus similarity-contingent neural adaptation can be time and cortical area dependent', *Journal of Neuroscience* **28**, 10631-10640.

Vogels, R. (2010), 'Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials', *Cerebral Cortex*.

Vogels, R. (2010), 'Mechanisms of visual perceptual learning in macaque visual cortex', *Topics in Cognitive Science* **2 (2)**, 239-250.

Vogels, R. (2009), 'Mechanisms of visual perceptual learning in macaque visual cortex', *Topics in Cognitive Science*.

Weinshall, D.; Hermansky, H.; Zweig, A.; Eshar, D.; Kombrink, S.; Ohl, F. W. & Pavel, M. (2010), 'Beyond Novelty Detection: Incongruent Events, when General and Specific Classifiers Disagree', *IEEE Transactions on pattern analysis and machine intelligence (PAMI)*.

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

Weinshall, D. & Zamir, L. (2007), Image Classification from Small Sample, with Distance Learning and Feature Selection, *in* '3rd International Symposium on Visual Computing', Springer Berlin / Heidelberg, , pp. 106--115.

Wetzel, W.; Ohl, F. & Scheich, H. (2008), 'Global vs. local processing of frequency-modulated tones in gerbils: an animal model of lateralized auditory cortex functions in humans', *Proc Natl Acad Sci* **105**, 6753-6758.

White, C.; Zweig, G.; Burget, L.; Schwarz, P. & Hermansky, H. (2008), Confidence Estimation, OOV Detection And Language Id Using Phone-To-Word Transduction And Phone-Level

Alignments, *in* 'IEEE Int. Conf. on Acoustics, Speech, and Signal Processing', pp. 4085--4088.

Willems, G.; Becker, J.; Tuytelaars, T. & Gool, L. v. (2009), Exemplar-based action recognition in video, *in* '20th British Machine Vision Conference (BMVC), London, UK'.

Witte, H.; Charpentier, M.; Mueller, M.; Voigt, T.; Deliano, M.; Garke, B.; Veit, P.; Hempel, T.; Diez, A.; Reiher, A.; Ohl, F.; Dadgar, A.; Christen, J. & Krost, A. (2008), 'Neuronal cells on GaN-based materials"Deutsche Physikalische Gesellschaft', Spring Meeting of the Deutsche Physikalische Gesellschaft, Berlin.

Zamir, L. & Weinshall, D. (2006), Feature Selection in Distance Learning from Small Sample, Used for Image Classification, *in* .

Zeghbib, A.; Fillbrandt, A. & Ohl, F. W. (2010), Inter-cortical networks synchronization under forcing auditory and visual stimuli, *in* '7th Forum of European Neuroscience (FENS)'.

Zeghbib, A.; Fillbrandt, A. & Ohl, F. W. (2010), Temporal coherent states evolution in multimodal learning, *in* '11th International Multisensory Research Forum (IMRF)'.

Zeghbib, A.; Fillbrandt, A. & Ohl, F. W. (2010), Behavioral relevance and physiological correlates of audiovisual interaction in rodent auditory and visual cortex, *in* 'DIRAC Workshop at ECML/PKDD 2010'.

Zeghbib, A.; Fillbrandt, A. & Ohl, F. W. (2009), Bidirectional information transfer between auditory and visual cortices, *in* '3rd International Conference on Auditory Cortex', pp. 158.

Zeghbib, A.; Fillbrandt, A. & Ohl, F. W. (2009), Alteration of brain states with high phase-coherence and transient states indicate the intermittency information processing in brain dynamics, *in* '7th Meeting of the German Neuroscience Society'.

Zeghbib, A.; Fillbrandt, A. & Ohl, F. W. (2009), Phase coherence evolution in cortical networks: adaptation to audiovisual stimulation with fixed inter-modality asynchrony, *in* 'International Multisensory Research Forum (IMRF)', pp. 391.

Zeghbib, A.; Fillbrandt, A. & Ohl, F. W. (2009), 'Phase change and Mutual Inter-cortical information flow under adaptation underlaying neuro-dynamical models', The XI. IfN-Forschungsseminars in Freyburg, Talk (Presentation).

Zeghbib, A.; Fillbrandt, A.; Deliano, M. & Ohl, F. (2008), 'Changes of oscillatory activity in the electrocorticogram from auditory cortex before and after adaptation to contingent, asynchronous audiovisual stimulation"International Multisensory Research Forum', IMRF, 9th Annual Meeting, p. 60.

Zeghbib, A.; Fillbrandt, A.; Deliano, M. & Ohl, F. (2008), 'Oscillatory activity in auditory cortex before and after entrainment with fast audiovisual stimulus sequences"Society for Neuroscience', Abstr. 851.5.

Zimmermann, K.; Matas, J. & Svoboda, T. (2009), 'Tracking by an Optimal Sequence of Linear Predictors', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(4), 677-692.

Zimmermann, K. (2008), 'Fast learnable methods for object tracking'.

Zimmermann, K.; Svoboda, T. & Matas, J. (2008), 'Simultaneous learning of motion and appearance"The 1st International Workshop on Machine Learning for Vision-based Motion Analysis', Marseille, In conjunction with ECCVΓўBЂB™08.

Zimmermann, K.; Svoboda, T. & Matas, J. (2007), Adaptive Parameter Optimization for Real-time Tracking, *in* .

Zimmermann, K.; Svoboda, T. & Matas, J. (2007), Adaptive Parameter Optimization for Real-time Tracking, *in* 'NRTL07, IEEE 11th International Conference on Computer Vision, 2007. (ICCV 2007)', pp. 1--8.

Zimmermann, K.; Svoboda, T. & Matas, J. (2006), Multiview 3D Tracking with an Incrementally Constructed 3D Model, *in* '3DPVT '06: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)', IEEE Computer Society, Washington, DC, USA, pp. 488--495.

Zweig, A. & Weinshall, D. (2007), Exploiting Object Hierarchy: Combining Models from Different Category Levels, *in* 'IEEE 11th International Conference on Computer Vision (ICCV 2007)'.

9.2. Deliverables

5th year

D6.11	Testing and validation plan	31.05.2010
D3.10	Report on single STS neurons to stimuli showing walking and running, as well as rare events, and qualitative comparison with conclusions from computational models for the same stimuli	30.06.2010
D4.14	Transfer of knowledge for learning of rare events	30.06.2010
D4.15	Adaptive learning of hierarchies	30.06.2010
D4.16	Current source density analysis of auditory cortex	30.06.2010
D6.12	Catalogue of basic scenes containing incongruent events	30.06.2010
D2.13	Acoustic stimuli with changing characteristics investigated with physiological and modeling approaches (OL)	30.09.2010
D2.14	Combining spike-based STRF measurements with analysis of the local field potential (LIN)	30.09.2010
D4.17	Multi-modal knowledge transfer for learning of rare events (IDIAP)	30.09.2010
D7.4	Fourth DIRAC summer workshop on cognitive engineering (KUL)	30.09.2010
D2.15	Incongruent acoustic scenes: Models, Detection and Identification (OL)	31.12.2010
D2.16	Final system for identifying unexpected acoustic inputs (BUT)	31.12.2010
D3.11	Qualitative comparison of neurophysiological findings and computation models for the same stimuli (KUL)	31.12.2010
D3.12	Fully functional implementation of a tracker tree pre-learned for a specific location and tested there, i.c. the OHSU Living Lab (ETHZ)	31.12.2010
D3.13	Concepts for self-learning tracker trees available as a report (ETHZ)	31.12.2010

D5.12	Generalized framework for rare event detection applied to realistic audio, visual and audio visual rare event detection problems (IDIAP) "	31.12.2010
D5.13	Current-source-density analysis of audiovisual interaction (LIN)	31.12.2010
D5.14	Measure of rareness of incongruent events (IDIAP)	31.12.2010
D5.15	Finalizing analyzes of the category-transfer experiment on the full data set (LIN)	31.12.2010
D6.13	Evaluation Results (FRA)	31.12.2010
D6.14	The DIRAC databases (FRA)	31.12.2010
D8.8	Annual Progress Report (OL)	31.12.2010

4th year

D6.6	First audio-visual data collection with AWEAR and OHSU system (FRA)	31.03.2009
D2.9	System for detection and description of out-of-vocabulary words in machine recognition of large vocabulary continuous speech (BUT)	30.06.2009
D2.10	STRF-based prediction of sound responses in auditory cortex (LIN)	30.06.2009
D2.11	Hierarchical acoustic scene classification scheme with detection of unexpected scenes (OL)	30.06.2009
D4.13	Modeling cortical cells with distance functions (HUJI)	30.06.2009
D6.7	Database of low level incongruencies (FRA)	30.06.2009
D7.3	Third DIRAC summer school on cognitive engineering (KUL)	31.08.2009
D1.8	Incongruence detection for detecting, removing and repairing incorrect functionality in low level processing (CTU)	31.12.2009
D2.12	Report on identification of repeatedly occurring out-of-vocabulary words (BUT)	31.12.2009
D3.8	Camera tracking and autocalibration for detecting and correcting camera de-calibration (CTU)	31.12.2009
D3.9	Incongruences detected between trackers working with weaker and stronger expectations about the world. (ETHZ)	31.12.2009
D5.11	Discriminative methods for multi-cue classification (IDIAP)	31.12.2009
D6.8	Database of incongruent human locomotion (FRA)	31.12.2009
D6.9	Database of recordings of scenario 1 (FRA)	31.12.2009
D6.10	Database of recordings of scenario 2 (OHSU)	31.12.2009
D8.6	Annual Progress Report (OL)	31.12.2009
D8.7	Update implementation plan (HUJI)	31.12.2009

3rd year

D1.6	Posterior Based and Spatiotemporal Features and on their Applications in Describing Audio-Visual Scenes (IDIAP)	30.06.2008
D2.7	Report on detection and description of unexpected out-of-vocabulary words (IDIAP)	30.06.2008
D2.8	Representations of Auditory Signals Derived by Biological Experiments and Statistical models (LIN)	30.06.2008
D3.6	Results on Biological and Computational Action Recognition	30.06.2008

	(ETHZ)	
D4.8	Adaptive Distance Learning and Classification (HUJI)	30.06.2008
D4.10	Learning-induced plasticity in cortical dynamics (LIN)	30.06.2008
D5.7	Report on Theoretical Analyzes, Simulation and Physiological Correlates in Rare (Unexpected Event Processing (IDIAP)	30.06.2008
D6.5	Technical Report on Detection of Audio-Visual Rare Events (31.12.2007 - Delayed M30) (IDIAP)	30.06.2008
D7.2	Second DIRAC Summer School on Cognitive Engineering (KUL)	30.08.2008
D1.7	Low level processing modules for AWEAR demonstrator (CTU)	31.12.2008
D3.7	Dynamic 3D Scene Analysis on Omni-Directional Video Data (ETHZ)	31.12.2008
D4.9	Neuronal Correlates of Categorization (LIN)	31.12.2008
D4.11	Information Transfer between Categories & Hierarchy Learning (HUJI)	31.12.2008
D4.12	Incremental Learning with Bounded Memory Growth (IDIAP)	31.12.2008
D5.8	Report on Neuronal Mechanisms Underlying Recalibration of Audiovisual Temporal-Order Judgment (LIN)	31.12.2008
D5.9	Report on Knowledge Transfer in Artificial and Living Cognitive Systems (IDIAP)	31.12.2008
D5.10	Report on High-Resolution Spatiotemporal Neuronal Dynamics from Auditory/Visual Cortex (LIN)	31.12.2008
D8.4	Annual Progress Report (OL)	31.12.2008
D8.5	Updated implementation plan (HUJI)	31.12.2008

2nd year

D1.4	Omnidirectional Image Acquisition and Processing Using High-Level Information (CTU)	30.12.2007
D1.5	Head-Geometry Beamformerwith Binaural Output and its Perceptual Quality Assessment (OL)	30.12.2007
D2.4	Results of Processing with Advanced Hierarchical Model Including Results on an Accepted ASR Task (IDIAP)	30.06.2007
D2.5	Acoustic Classification Method (OL)	30.06.2007
D2.6	Detection of Unexpected Words In Machine Recognition of Speech (IDIAP)	30.06.2007
D3.3	Conclusions from the First Neurophysiological STP Single-Cell Experiments (KUL)	30.06.2007
D3.4	Biological Motion Detection Based on the Learned, Statistical Representations (ETHZ)	30.06.2007
D4.4	Test & evaluation of motion boundary detection algorithm (HUIJ)	30.06.2007
D4.5	Conclude Categorization Study with PEC/NEC, Design Developmental Study, Feature Category Relationship (HUJI)	30.06.2007
D4.6	Categorization of Level and/or Temporal Dynamics (LIN)	30.06.2007
D5.4	Report on High Level (Categorical) Audio-Visual Integration Results (LIN)	30.06.2007
D5.5	Example/Demo of Hierarchical Multimodal Fusion (IDIAP)	30.06.2007
D6.1	Application Scenario (Updated version) (IDIAP)	30.06.2007
D6.2	Omnidirectional Camera Tracking (CTU)	30.06.2007

D6.3	Annotated Bibliography Webpage (IDIAP)	30.06.2007
D6.4	Website for Audio-Visual Data (IDIAP)	30.06.2007
D6.5	Technical Report on Detection of Audio-Visual Rare Events (IDIAP)	31.12.2007 Delayed M30
D7.1	First DIRAC Summer School on Cognitive Engineering (KUL) and its Updated version	30.06.2007 Update: 02.10.2007
D3.5	Dynamic 3D Scene Analysis Using Cognitive Loops (KUL)	31.12.2007
D4.7	Object Detection from Small Sample, Optimization (HUJI)	31.12.2007
D5.6	Report on electrophysiological correlates of audio-visual temporal order judgments (LIN)	31.12.2007
D8.2	Annual Progress Report Y2 (IDIAP)	31.12.2007
D8.3	Updated Implementation Plan (IDIAP)	31.12.2007

1st year

M7.1	Bundled notes of the Ice Breaker Workshop (KUL)	09.05.2006
M7.2	Establishment of Internship Programs (undergraduate & graduate) (KUL)	30.06.2006
M7.3	Establishment of Liaison with other (European) Consortia - list of contact person(KUL)	30.06.2006
D2.1	Data Recorded from Different Acoustic Environments and its Updated Version (31.12.2006) (OL)	30.09.2006
D1.1	Omni-directional Sensor, and Features for Tracking and 3D Reconstruction(CTU) , Prototype	30.12.2006
D1.2	Feature Detectors for Body Parts, Tracking, and Fast Object Detection(CTU) , Prototype	30.12.2006
D1.3	Acoustic Features that Employ Limited Frequency Ranges and Longer Temporal Spans (IDIAP)	30.12.2006
D2.2	Recordings of Spectro-Temporal Receptive Fields (STRFs) from Gerbil Auditory Cortex(LIN)	30.12.2006
D2.3	Features of Audio Signals, Obtained with Different Modeling Techniques (IDIAP)	30.12.2006
D3.1	Framework for Bottom-Up 3D Reconstruction (KUL) , Prototype	30.12.2006
D3.2	Setup and Stimuli for Neurophysiological Experiments (KUL) , Prototype	30.12.2006
D4.1	Evaluation of Localization Algorithm and Retrieval) (HUIJ)	30.12.2006
D4.2	Prototype of the VR Environment (HUIJ)	30.12.2006
D4.3	Learning-induced plasticity in cortical receptive fields) (LIN)	30.12.2006
D5.1	Setup of experimental paradigm for studying audio-visual fusion in rodent cortex (LIN)	30.12.2006
D5.2	Combination of Nonlinear Classsification(IDIAP)	30.12.2006
D5.3	Framework for Cognition-Based Fusion (OHSU)	30.12.2006
D6.1	Application Scenarios (ALL)	30.12.2006
D8.0	Periodic Progress Report Y1(IDIAP)	30.12.2006
D8.1	Updated Technical Annex (IDIAP)	30.12.2006