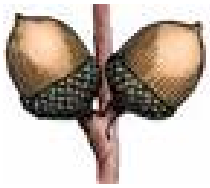

Project no. 034362

ACORNS

Acquisition of COmmunication and RecogNition Skills

Instrument: STREP
Thematic Priority: IST/FET

Periodic Activity Report

Period covered: from 1 December 2006 to 30 November 2007

Date of preparation: 24 December 2007

Start date of project: 1 December 2006 Duration: 36 months

Project coordinator name: Prof. Lou Boves
Project coordinator organisation name: Radboud University, Nijmegen
Revision [1]

1. Executive summary

ACORNS, Acquisition of COmmunication and RecogNition Skills
www.acorns-project.org

Participants:

- Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands (co-ordinator)
- Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland
- Sound and Image Processing Laboratory, Royal Institute of Technology, Stockholm, Sweden
- Speech and Hearing Research Group, University of Sheffield, United Kingdom
- Center for Processing Speech and Images, Katholieke Universiteit Leuven, Belgium

Coordinator contact details: Dr. Lou Boves, CLST, P.O.Box 9103, 6500 HD Nijmegen, The Netherlands, l.boves@let.ru.nl; Tel: + 31243612902.

1.1 Summary description of project objectives

ACORNS intends to develop and test a complete computational model of the memory-prediction theory of intelligent behaviour applied to language acquisition and speech communication. The input for the model will consist of audio signals in combination with symbolic representations of the environmental context to which spoken utterances refer. Such a model is by necessity modular in structure. In accordance with the modular structure of the model, the project work plan is structured in five technical work packages, four of which are devoted to producing a specific processing module, and a fifth one dedicated to the integration of the modules and conducting the experiments that will prove the capability of the model to account for the acquisition of language and communication skills.

The **front-end processing** module, to be developed in WP1, will result in a rich representation of audio signals that is suitable to characterise and process essentially all ecologically relevant sounds and to model different sources independently, with a strong focus on features that are important for speech processing.
The **pattern discovery** module, which is the focus of WP2, will design and implement computational models that can detect recurring patterns in input signals that can be linked to other memory representations, including motor actions. A novel approach is being investigated to representing and patterning speech and audio signals. Speech is represented in the form of codes for short (0.5 to 2 ms) segments of the audio signal, and a non-linear process is being developed for mapping signal fragments to a permutation space, based on an appropriate metric.
Memory organisation and access is the focus of WP3. We will develop computational representations of the different types of memories that are implied in the memory-prediction theory and other theories of the neural storage and processing of speech signals. An important aspect of memory processes is how representations of novel patterns can form and be stored.
In WP4, **Information discovery and integration,** three approaches will be investigated to address associative memories. The first approach takes its guidance from content addressable memories (CAM), especially the form of CAM that handles fuzzy and incomplete codes for addressing the contents of the memory. The second approach borrows from Latent Semantic Analysis, developed for document retrieval, but also proposed as a model for speech understanding (i.e., comprehending the semantic content of spoken

utterances). The third approach is based on the assumption that closely related patterns link to each other, so that it is possible to quickly identify all patterns that resemble some input to be recognised.
For WP5, **Interaction and communication**, the aim is to integrate the processing modules in an integrated system that simulates speech acquisition. The system or agent, dubbed Little Acorns, is endowed with the intention to learn a continuously growing vocabulary in order to maximise the positive feedback from its environment. Three increasingly more complex stages, one for each year of the project, are defined. In the first stage, Little Acorns must learn to pay attention to sounds in its environment, and more specifically to speech produced by one or two 'parents'. At the end of the first phase our artificial infant must have acquired the basic skills needed to understand that it is being addressed. In the second phase, Little Acorns must learn a basic vocabulary by listening to simple speech utterances that will be presented in the context of corresponding objects, actions and concepts. In the third and last phase of language acquisition that we will simulate, Little Acorns will learn a vocabulary of some 250 words, including verbs for actions that can reasonably be performed on the objects that correspond to nouns and adjectives (e.g., colour names, size (big, small), shape (square, round), etc.). Thus, at the end of this development phase Little Acorns should be able respond with references to more than one 'concept' in its memory, plus some reference to a relation between those concepts. The autonomous learning skills of Little Acorns will be demonstrated by introducing new objects into the world, for which she will be able to learn new words.

1.2 Work performed

During the first year baseline modules were developed and delivered for **front-end processing** and **information discovery** that were **integrated in a system** that allows us to conduct experiments to simulate language acquisition. For four languages (Dutch, Swedish, Finnish and English) a small corpus of infant directed and adult directed speech has been recorded for use in experiments with language acquisition. The integrated system and the corpora have been used to perform a set of baseline experiments that have shown that a properly designed computational model can learn to distinguish between utterances that refer to different objects on the basis of very general learning principles. Linguistic representations are emergent properties of the model, rather than information that is built into the model before the learning starts.
The processing and internal representations of the model can be mapped on a general model of speech processing that is compatible with the memory-prediction theory and at the same time reflects the results of a large body of psycholinguistic and psychological experiments. This memory and processing model is shown schematically in Fig. 1.
Substantial progress has been made in developing a novel approach to signal-driven structure discovery in WP2. For the moment, this has resulted in a patent application and operational software that is ready for integration in the system for conducting language acquisition experiments.

A very important dissemination activity took place at the end of the first year of ACORNS. Co-funded by the European Science Foundation (ESF) and the Netherlands Organisation for Scientific Research (NWO) an on-invitation only workshop was organised. This three day workshop with the title "Models of language evolution, acquisition, and processing" was held in Leuven, Belgium. The aim of the workshop was to present the results of the first year of ACORNS to the Scientific Advisory Board, and to define a roadmap for cognitively and evolutionary inspired research on speech acquisition and (automatic) speech processing. The outcome of the workshop will be a published book containing chapters based on presentations given by participants of the workshop.

1.3 Results achieved

In the first year we have constructed operational modules for acoustic pre-processing, for information discovery and for conducting language acquisition experiments. In addition, we have created an initial version of a learning agent (an artificial baby, as it were) that, in interaction with a caregiver (also artificial) can learn language by processing multimodal input. The eventual memory architecture of the learner is shown in Figure 1. Quite naturally, the implementation created in the first year of the project is somewhat simpler. To make possible a form of reinforcement learning, the learning agent indicates what he has understood, and the caregiver gives feedback about his/her satisfaction with the learner's performance. The set-up of the learning experiments is shown in Figure 2. In order to be able to conduct biologically and

cognitive learning experiments we have recorded small corpora of infant directed and adult directed speech in four languages (Dutch, English, Finnish and Swedish) and used the recordings for conducting experiments. The results of the experiments conducted during the last two months of the first year used only one approach to pattern and information discovery, based on Non-negative Matrix Factorisation, a technique reminiscent of Latent Semantic Analysis. The results of the experiments show that a learning agent can discover structure in input speech and relate that to some 10 different objects.

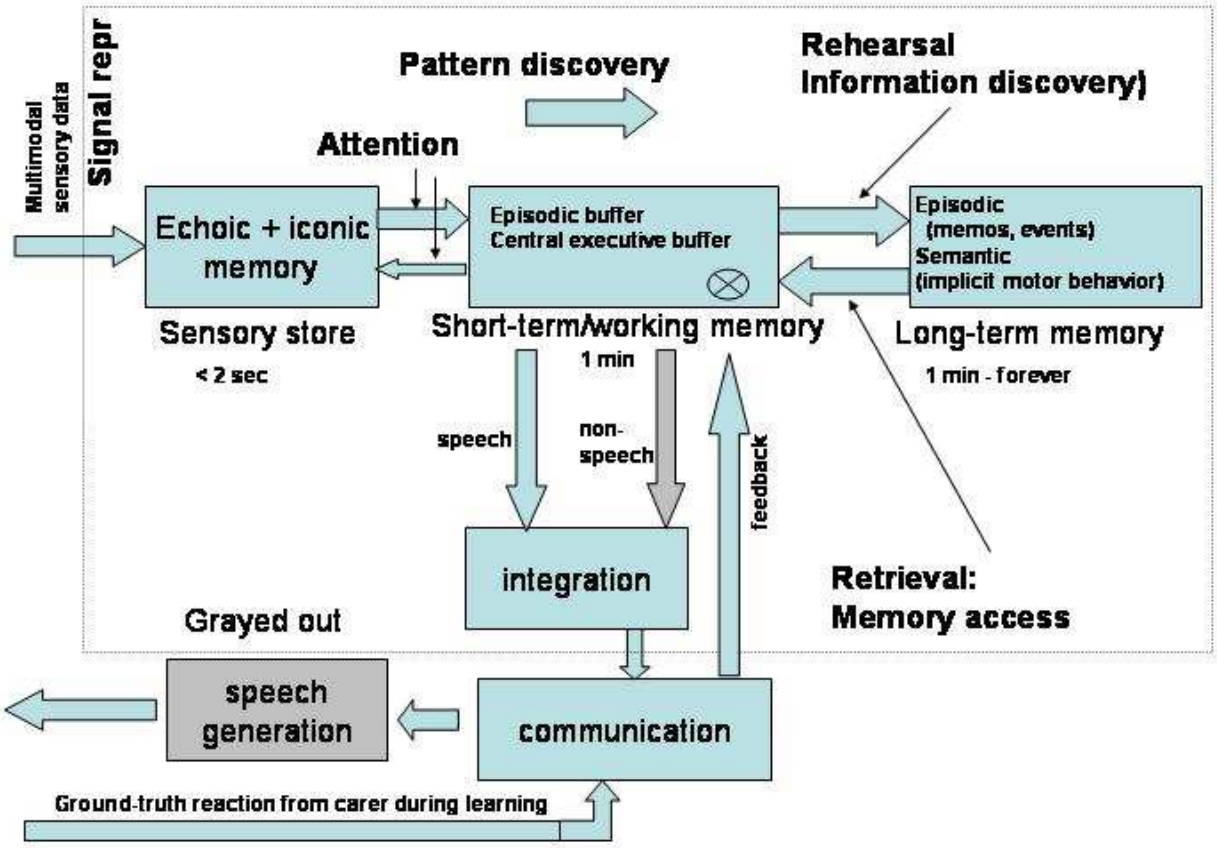

Figure 1 Schematic overview of the memory and processing architecture in ACORNS.

1.4 Expected end-results

In the first year of the project we intended to show that an artificial agent is able to discover structure in child directed speech signals, build internal representations of the acoustic signals in memory, and link the acoustic representation to a small number of physical objects that are (virtually) present in the scene while a corresponding utterance is spoken. These targets have been reached, despite the fact that some partners were only able to put together a complete team in the second half of the year. Thanks to intensive and effective collaboration between the partners we have been able to build a platform for conducting learning experiments, to integrate the module for conventional feature extraction from WP1 and a module for information discovery from WP4 in the platform, and conduct experiments. The design and results of the experiments show how the pattern discovery techniques under development in WP2 can be integrated in the platform now that initial software modules are available. It has also become clear how the results of the experiments performed so far can be mapped onto the memory architecture under development in WP3.

The focus of the research in the next year will be on the development of novel features in the acoustic pre-processing, further development of techniques for structure discovery and information discovery, and on a tighter integration of these modules in the platform for conduction learning and interaction experiments. The learning task will be more complicated, in that the artificial agent must be able to handle a larger number

(±50) of concepts (not only nouns, but also verbs and adjectives), to discover multiple semantic units in an utterance and to build internal representations that can be linked both to the acoustic signals and the semantic/pragmatic value of an utterance and can be used as a stepping stone for learning additional words and concepts. We will also investigate whether the emergent internal representations can explain the loss of sensitivity to non-native phonetic contrasts that is observed in babies after their first birthday. Additional success criteria will be defined in the first project meeting, which is will take place in March 2008.

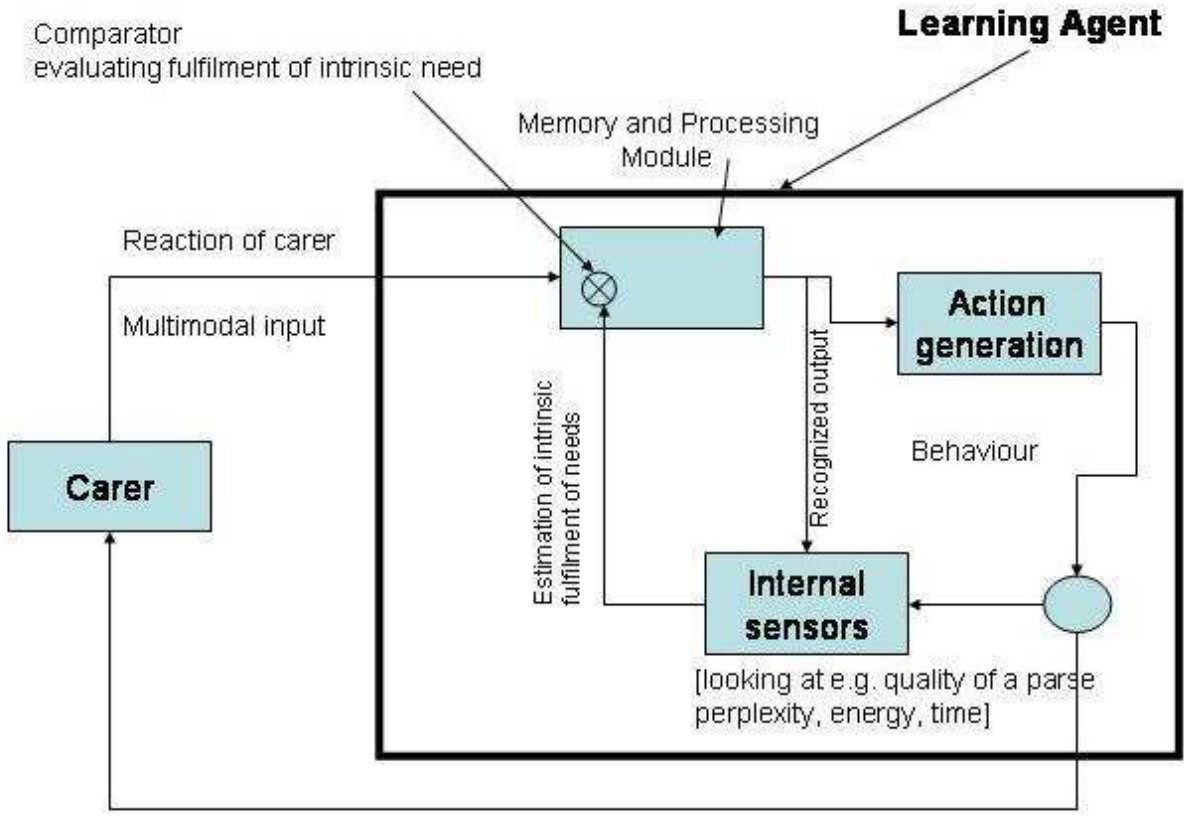

Figure 2 Schematic overview of the learning process.

1.5 Intentions for use and impact

ACORNS was instrumental in the organisation of an ESF Exploratory workshop entitled "Models of Language Evolution, Acquisition and Processing". One of the aims of this workshop was to investigate options for future research in the general field addressed by the ACORNS project, as well as options for organising the field. It is expected that this workshop will indeed result in proposals for follow-up projects and for a network of researchers in the field of computational modelling of language acquisition and processing. The workshop will result in a book, to be puiblished by a major publisher. It is expected that this book will have a considerable impact in the field.

TKK filed apatent application for a novel procedure for bottom-up segmentation of speech signals. The procedure can also be applied to non-speech time signals. The procedure is based on discovering local structure in a time signal, and subsequent clustering of local structure elements to form a number of classes or categories that may carry information about the process that generated the signals in the first place.

At the end of the project ACORNS will release public domain software supporting computational modelling of language acquisition and processing. Preliminary versions of part of the prospective software package have been exchanged between the partners for test and for use in conducting modelling experiments.

Already in its first year ACORNS has resulted in a number of conference papers, one journal paper, and several master theses.

2 Project objectives and major achievements during the reporting period

ACORNS intends to develop and test a complete computational model of the memory-prediction theory of intelligent behaviour as it applies to language acquisition and speech communication. The input for the model will consist of audio signals in combination with symbolic representations of the environmental context to which the audio signals refer. Such a model is by necessity modular in structure, and individual modules will be designed, implemented and tested by the partners. In accordance with the modular structure of the model that we intend to create and test, the project work plan is structured in five technical work packages, four of which are devoted to producing a specific processing module, while WP5 is responsible for the integration of the modules and for conducting the experiments that will prove the feasibility of the memory-prediction theory to account for the acquisition of language and communication skills. The following four modules will be developed: WP1 front-end processing, WP2 pattern discovery, WP3 memory organisation and access, WP4 information discovery and integration, and WP5 is responsible for interaction and communication.

The **front-end processing** module should result in a rich internal representation that is suitable to characterise and process essentially all ecologically relevant sounds and to model different sources independently.
The **pattern discovery** module will design and implement computational models that can detect recurring patterns in the input signals and that can be linked to other memory representations, including motor actions. The module will initialise and bias the learning system so that it is optimally suited to handle important signal characteristics. A novel approach will be investigated to representing and patterning speech and audio signals under development by the Helsinki University of Technology. Here, speech is represented in the form of codes for short (0.5 to 2 ms) segments of the audio signal, known as Discrete Model Elements (DME). This method uses a nonlinear process for mapping signal fragments to a permutation space, based on an appropriate metric. The method dramatically reduces the computational load when compared to classical linear mappings and transformations.
Memory organisation and access will focus on the development of suitable computational representations of the different types of memories and the processing that takes place. An important aspect of memory processes is how representations of novel patterns can form and be stored.
For **Information discovery and integration** three approaches will be investigated to address associative memories. The first approach takes its guidance from content addressable memories (CAM), especially the form of CAM that handles fuzzy and incomplete codes for addressing the contents of the memory. The second approach borrows from Latent Semantic Analysis, developed for document retrieval, but also proposed as a model for speech understanding (i.e., comprehending the semantic content of spoken utterances). The third approach is based on the assumption that closely related patterns link to each other, so that it is possible to quickly identify all patterns that resemble some input to be recognised.
For the **Interaction and communication** module, the aim is to integrate all processing modules sketched above in an integrated system that simulates speech acquisition. The system, Little Acorns, is endowed with the intention to learn a continuously growing vocabulary in order to maximise the positive feedback from its environment. Three increasingly more complex stages, one for each year of the project, were defined for Little Acorns to learn speech. In the first stage, little Acorns must learn to pay attention to sounds in its environment, and more specifically to speech produced by one or two 'parents'. At the end of the first phase our artificial infant must have acquired the basic skills needed to understand that it is being addressed, and show this by the virtual equivalent the head turning paradigm used in experiments with very young children. In the second phase In the second phase, Little Acorns must learn a basic vocabulary by listening to simple speech utterances that will be presented in the context of corresponding objects, actions and concepts. The correspondence will be probabilistic, but in this phase of the process the probabilities will be relatively high, so as to help Little Acorns to learn appropriate relations and virtual responses. In the third and last phase of language acquisition that we will simulate, Little Acorns will learn a vocabulary of some 250 words,

including verbs for actions that can reasonably be performed on the objects that correspond to nouns and adjectives (e.g., colour names, size (big, small), shape (square, round), etc.). Thus, at the end of this development phase Little Acorns should be able respond with references to more than one 'concept' in its memory, plus some reference to a relation between those concepts. The autonomous learning skills of Little Acorns will be demonstrated by introducing new objects into the world, for which she will be able to learn new words.

Objectives for the reporting period were:

The main objective for the reporting period were:
Deliver base-line versions for the separate modules and to design the outline for the memory representations to be used in ACORNS. The modules had to be integrated into an overall architecture. Experiments with real input speech (parentese and normal speech) should demonstrate that the ACORNS agent is able to learn ten words from scratch.

At the end of the first year the system must be able to build internal representations of some 10 words which are not too difficult to distinguish acoustically and which will be produced by four speakers. However, it must be able to handle somewhat similar words like 'papa' and 'mamma'. The system must be able to form and access these representations from continuous speech input, be it that the input utterances will have the characteristics of 'parentese', i.e. the somewhat exaggerated type of speech that is often used to address babies. The system's capability to learn and understand words will be demonstrated in several (closely related) ways. First of all, when presented with suitable novel inputs (utterances containing words that the system has learned, and spoken by a person with a voice similar to one of the voices it has learned, in an appropriate visual context) the system must be able to respond appropriately, i.e., it must be able to select the response associated with the word. Second, in order to prove the capability for learning words, the system must be able to segment the acoustic input in such a manner that the starting and ending points of the relevant words are properly indicated. Thirdly, we will analyse the internal representations of the input speech (utterances, words, and sub-word units) to verify that they are suitable for subsequent learning of ever more words on the basis of some kind of acoustic phonetic analysis of the speech input. If the system has learned several words that contain the same speech sounds, the representations should show sufficient commonalities.

We will demonstrate that the learning capability of the system is language independent by using several different languages in combination with the same visual inputs. Also, the resulting communicative behaviour of the learning system should not be dependent on the order in which training utterances are offered. For all languages comparable learning performance will be demonstrated. We will allow for some degree of variation in the internal representations of utterances, words and sub-word units as a function of the language the system is acquiring and the voice characteristics of the speakers who address the system. However, the representations should be independent of the order of presentation of the training speech.

Results

As can be seen from the detailed accounts of the work performed in the individual work packages in chapter 3 of this report, it is fait to say that ACORNS has reached all targets set for the first year. Most specifically, we have succeeded to show that an artificial agent is able to learn a number of words, using a simple and general learning method, from real acoustic and simulated visual input only.

Radboud University Nijmegen submitted a proposal to the Netherlands Organisation for Scientific Research NWO for a research programme (one post-doc and two PhD students) that is closely related to ACORNS. This proposal was accepted. The newly acquired research programme will start in the spring of 2008. There will be close collaboration and cross-fertilisation between the two projects.

3 Workpackage progress of the period

3.0 Workpackage 0 Project Management

The Workspackage Management is divided into two Tasks: Scientific Management and Financial and Administrative Management.

The tasks for this reporting period were:

- Manage the project scientifically
- Install the Scientific Advisory Board
- Prepare, conduct, and report on meetings
- Deliver Consortium Agreement
- Make the project Representation for the Web site (www.acorns-project.org)
- Deliver Periodic Activity Report, and Management Report
- Take care of the financial issues in the project

Achievements:

- The project was well management. All deadlines were met, there is very good collaboration between the partners and the planned direction is followed.
- The Scientific Advisory board was installed. Most members were invited to the workshop that was organised together with funding of ESF and NWO. At this meeting a special presentation was dedicated to ACORNS so that the SAC members could get a good picture of the achievements of the project in the first year. Useful feedback was received during the discussions. These relate e.g. to the status of the tags that are associated with the audio input, the assumptions that are made in the cognitive model, and the speech database.
- Five fruitful meetings took place and three conference calls were held.
- The Consortium Agreement was delivered.
- The project Presentation was made for the ACORNS website.
- The Activity, and management reports were delivered.
- Els den Os visited Mr. Paul Hearn to discuss the project.

Table 3.0.1: Deliverables List

Del. no.	Deliverable name	Workpackage no.	Date due	Actual/Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months *)	Lead contractor
D0.1	Project Presentation	0	M6	M6	1	1/2	RUN
D0.2	Activity Report	0	M12	M12	3	1/2	RUN
D0.3	Dissemination and Use	0	M12	M12	2	1/3	RUN
D0.4	Management Report	0	M12	M12	1	1/3	RUN

***) if available**

- List of milestones, including due date and actual/foreseen achievement date

For the milestones, see the deliverables.

3.1 WP1 Signal Representations

1. Workpackage objectives from the TA:

The objective of WP1 is to develop information a rich representation of speech and sound by discovery and integration mechanisms by defining a compact representation based on models of human perception and by extracting signal features that are known to be relevant for human speech and sound. These objectives are still relevant. The objective for this period was to create modules for a conventional feature set and to prepare the auditory models to be used in the sensitivity analysis.

2. Progress towards objectives

We delivered the set of standard features that are commonly used in speech recognition and are based on prior experience and knowledge according to the milestone and deliverable schedule. The standard ACORNS features are: logarithmic mel spectrum, mel-frequency cepstral coefficients (MFCCs) and the pitch period track. The basic feature set is augmented with delta and delta-delta versions of the MFCCs. The features are already in use as an initial basis for the work performed in other work packages in ACORNS. In addition, these features will be used in WP1 as a reference for the proposed feature selection method as it described in ACORNS Technical Annex.

We carefully tested our implementation of the standard features. As the MFCCs are commonly used we compared our implementation with available implementations. We found that the implementations available from various sources are similar but not equivalent. We compared with implementations available in the VOICEBOX toolbox, AUDITORY toolbox, in RASTAMAT and in HTK. The results of the comparison are described in deliverable D1.1.

We perform pitch estimation with a method that is essentially that of the ITU G.729.1 speech coder. In comparative testing, we have found that this estimation procedure has the lowest sensitivity to pitch doubling and pitch halving despite the fact that it has low computational complexity and facilitates frame-by-frame processing.

As required by milestone 1.1, KTH prepared software that describes two auditory models: a spectral model developed by van de Par et al. at Philips and the PEMO auditory model developed at Oldenburg. The PEMO model is partly based on software obtained from a third party.

3. Deviations form the workplan

There were no deviations from the workplan.

4. Connection to other work packages

The modules delivered under WP1 are already in use in WP4.

Table 3.1.1: Deliverables List for WP1

Del. no.	Deliverable name	Work package no.	Date due	Actual/Forecast delivery date	Estimated indicative person-months	Used indicative person-months	Lead
WP1							
D1.1	Modules for conventional feature set	1	M12	12	15	10	KTH
D1.2	Modules for a) augmentation of standard spectral features with a stream of milli-second and decisecond features and evaluation on specific phone classification tasks and b) feature selected by sensitivity-analysis method	1	M24	17	17	2	KTH
D1.3	Final Modules for features derived with sensitivity-analysis method criterion, with quantitative evaluation	1	M36	15	15	0	KTH

Table 3.1.2: Milestones List for WP1

Milestone no.	Milestone name	Workpackage no.	Date due	Actual/Forecast delivery date	Lead contractor
M1.1	Conventional feature set completed	1	M6	M9	KTH
M1.2	Auditory models for sensitivity analysis completed	1	M12	M12	KTH
M1.3	Validation of sensitivity analysis method and method based on augmentation with milli-second and decisecond features	1	M18	M18	KTH
M1.4	New features based on sensitivity analysis method	1	M30		KTH

3.3 WP2: Signal Patterning (M1-M12)

3.3.1. Workpackage objectives from the Technical Annex. Two tasks:

1.1 Task 1 - Pattern discovery using discrete model elements (DME)

- Investigate the emergence of basic patterns and the ways in which these can subsequently be used to represent more complex units.
- Consider the speech signal as a sequence of very short acoustic phenomena that are treated as 'data' without reference to phonetics or linguistics.
- Apply metrics in the articulatory-acoustic space.
- Implement a segmentation strategy based on attendant measures of constancy utilising classical linear signal processing as well as scale-invariant methods to form DMEs.
- Channel results of segmentation towards the identification and labelling of clusters.
- Five (5) subtasks:

Subtask WP2.T1.1
Build a Pattern Discoverer (PD) that produces unsupervised segmentation and descriptions of segments found. Preliminary classification of segments. ➔ Milestone M2.1.1A
Study auditory type pre-processing combined with the DME concept. ➔ Milestone M2.1.1B

Subtask WP2.T1.2
Adapt PD module to provide both time and auditory domain segmental measures enabling its use for higher level processing. ➔ Milestone M2.2.1

Subtask WP2.T1.3
Define compact coding representations for higher-level processes enabling the generation of higher-order representations. These will be used for clustering. ➔ Milestone M2.1.3

Subtask WP2.T1.4
Analyse temporal structures in wider temporal windows by using chains of linked DMEs. Study how to represent auditory stimuli and memory. ➔ Milestone M2.1.4

Subtask WP2.T1.5
Build a PD equipped with self-directed search. Define a set of segmental quality measures. ➔ Milestone M2.1.5

1.2 Task 2 - Pattern discovery with computational mechanics approaches

- Investigate the applicability of pattern discovery techniques to discover the essential structures and patterns in speech signals.
- Three (3) subtasks:

Subtask WP2.T2.1
Apply a CMM-based paradigm to the problem of discovering patterns in signal representations generated by the two approaches in WP1. Develop CMM theory and tools to enable incremental learning, i.e., a type of learning in which the trade-off between stochastic complexity of the input data and previously acquired knowledge, changes as the learning process proceeds. ➔ Milestone ~~M2.2.1~~ M2.1.2 *(M2.2.1 is a typographical error in the TA; it should read M2.1.2).*

Subtask WP2.T2.2
Generalise the learning to discovering structure and patterns in sequences of elementary units as a means for learning words and multiword expressions, including other hierarchical structure such as segments, diphones, and syllables. ➔ Milestone M2.2.2A

Apply CMM for discovering meaningful structure on the syntactic level. ➔ Milestone M2.2.2B

Subtask WP2.T2.3
Report on the full integration of CMM learning to ACORNS. ➔ Milestone M2.2.3

2. Progress towards objectives

Task 1 - Pattern discovery using discrete model elements (DME)
From M1 to M6 work was carried out by WP2 on the theoretical design of a bottom-up methodology utilising DME that could be used for segmenting speech signals. In M7 the implementation of a DME-based system was commenced that is based on fine scale analysis of the speech spectrum. By M12 a blind-segmentation system had been written in MATLAB that could reliably detect changes in speech signals that correlated closely with acoustic-phonetic boundaries that humans perceive. Large speech corpora, e.g., TIMIT and TI-DIGITS among others, were used to explore the characteristics of the bottom-up approach to segmentation. The corpora were also used to determine the levels of blind bottom-up segmentation performance obtained and compared to what currently exists in literature. The developed system performs classification of the detected segments and provides reliability indices for them.

One of the first year goals of ACORNS was to be able to recognise a small set of keywords, e.g., 10 words, in continuously spoken speech (cf. TA, pg. 44). For this reason, a proof-of-concept experiment was conducted with the developed bottom-up segmentation and clustering system developed in WP2 that indicated that the units the system was able to detect and cluster could successfully be used to determine similar clustered sequences of acoustic phenomena arising from different speakers, e.g., units such as words could be readily detected and identified. The details of these experiments and their results are available in WP2's deliverables.

Task 1 - Deliverables
From the TA: *"We will report on our work regarding the PD and DME modules defined in subtask WP2.T1.1 and includes milestones M2.1.1A and M2.1.1B."*

Work regarding the PD and DME modules has been reported through:

- the Pattern-Discovery software module (including several software updates) that has been provided by WP2 to the ACORNS consortium.
- In M10 an abridged report (8 pages) entitled "WP2: Plan, Strategy, Methods, Integration, and Current Status" was distributed to ACORNS at the group's Stockholm meeting. The full report (ca. 15 pages) has been made available to ACORNS by the end of M12.
- In M11 a patent application (FIN-20075696) was submitted covering the methods described in the above-mentioned bottom-up segmentation system.
- Weekly internal WP2 reports have culminated in the M.Sc. thesis publication (94 pages) of Okko Räsänen (WP2's PhD student) during M12. This thesis has already been made available to ACORNS and further dissemination of the work to appropriate journals is in progress.

Task 2 - Pattern discovery with computational mechanics approaches
From M4 to M6 preliminary work was carried out by Gustav Henter of KTH on a literature survey of the computational mechanics models (CMM) approach and its applicability to the problems faced by ACORNS in pattern discovery. This theoretical review was presented in M7 to the ACORNS group. Since the potential offered by a CMM algorithm entitled *Causal State Splitting Reconstruction* (CSSR) seemed applicable, an available implementation of CSSR was tested by Henter on a lengthy sequence of one million discrete symbols that was generated by Bosch, (Nijmegen) and introduced a deliberate noise (error) component.

On October 3, 2007 a one-day workshop concerning CMM was organised by KTH in Stockholm where members from Nijmegen, Stockholm, and Helsinki were present. Results from the tests designed in M7 and

that were carried out by Henter were presented. Results on long sequences corrupted with noise indicated that:

a) CSSR was very sensitive to noise and in its current form could not offer any substantial/practical solutions to pattern discovery without modifying parts of the algorithm and perform further testing,
b) the available CSSR algorithm used in the tests (implemented in C by K. L. Shalizi) contained errors. This may have (or may not have) influenced the results obtained in a).

At the workshop it was decided that CSSR still holds potential since, e.g., it had been successfully applied to linguistically related tasks. However, it was decided that the algorithm should be re-implemented and at the same time debugged. Furthermore, modifications should also be made to it so that its performance in noise could be improved. Currently WP2 is re-implementing the CSSR algorithm of Shalizi in another environment so that modifications and further testing of CMM applicability to pattern discovery in non-ideal noisy conditions can be evaluated more readily.

Task 2 - Deliverables

From the TA: *"We will report on the design and implementation of a comprehensive operational environment for CMM learning (subtask WP2.T2.1) and includes milestone M2.2.1."*

Work regarding CMM and CSSR has been reported through:

- A presentation and report (available on the ACORNS wiki) regarding CMM and CSSR to ACORNS by Henter (KTH) was presented at the group's June, 2007 meeting (M7).
- A one-day (pre-quarterly meeting) workshop was held on Oct. 3, 2007, at Stockholm. KTH, TKK, and RUNijmegen participants were present.
- A second presentation and report (available on the ACORNS wiki) dealing with the influence of noise on CMM learning was presented by Henter (KTH) in M10.
- A literature report on CSSR was issued in M10 by Henter (KTH). This report will be made available as a PDF document on the ACORNS wiki in M12.

A comprehensive operational environment for CMM learning is not available at this moment yet; we are currently still working on an independent implementation of Shalizi's algorithm. This will allow new experiments to study measures to handle noise.

3. Deviations from the project work programme

Task 1:

The acquisition of a suitable PhD student occurred only in M7 and thus made meeting the projected Milestone M2.1.1A completion date of M6 not possible.

Milestone M2.1.1B (auditory type pre-processing combined with the DME concept) has not yet been pursued since segmentation performance, using linear spectral representations in WP2's developed segmentation algorithm, has exceeded what is currently described in literature. For this reason, further effort has been channelled to keep on improving the performance of the current system since above-expected performance has been obtained. Another reason for moving the aspects of auditory and temporal research forward was since WP1 features became available only in M10, as did most of the other WP's software packages. Since auditory type pre-processing may well offer further benefits, experiments that will study its use in both the segmentation and clustering processes, have been planned and scheduled to be ready for M15 (milestone M2.1.1B).

However, in other areas of Task 1, WP2 is ahead of schedule. Progress on second year subtask WP2.T1.3 and third year subtask WP2.T1.5 has already been made. Part of the Year 2 goals of *defining compact coding representations for higher-level processes, including statistical, on-the-fly analyses of DME elements themselves, enabling the generation of higher-order representations*, has already been accomplished up to M2.1.3. Preliminary results of these studies are reported in "*Speech Segmentation and Clustering Methods*

for a New Speech Recognition Architecture" (M.Sc. thesis of O. Räsänen, 2007) in sections dealing with clustering. Additionally, part of the third year goals of *defining a set of segmental quality measures* (part of M2.1.5) has been addressed and reported in the previous mentioned thesis. Finally, the third year emphasis on *noise robust methods* (please see Technical Annex, page 35) has been covered by studying the effect of noise on segmentation, also reported in the previous reference. Fulfilling M2.1.3 (scheduled for M24) and partially M2.1.5 (scheduled for M30) already at M12 was a choice made for reasons of efficiency since the line of work related to blind-segmentation and clustering tied together in a natural manner.

Three minor changes are being requested by WP2 to be made to the text found in WP2 Task 1 Subtasks 1, 2, 3, and 5, in the ACORNS Technical Annex (found on page 33):

1. Subtask WP2.T1.1: Top of page 33: The sentence starting with:

"Halfway through the project the PE-module …" should read "Halfway through the project the PD-module …"

Also, this sentence, as well as all text following it existing under Subtask WP2.T1.1 should actually exist in Subtask WP2.T1.2. Now Subtask WP2.T1.2 reads.

> Subtask WP2.T1.2
>
> *Halfway through the project the PD-module will provide both time-domain and auditory domain segmental measures for higher level processing. It is noted that the auditory spectral representation will initially use a short time window; wide temporal representations in terms of combinations of short-time elements are provided in milestone subtask WP2.T1.4.* The second sub-task is devoted to adapting the PD module so that it provides both time-domain and auditory domain segmental measures enabling its use for higher level processing (M2.1.2).

This moving forward of these two sentences of text is reasonable since "Halfway through the project" means M18, the time M2.1.2 is due (ACORNS is a 36 month long project, refer also to pages 48 and 59 of the Technical Annex where the correct M18 due dates are already listed).

2. Subtask WP2.T1.2: The reference to Milestone M2.2.1 is a typographical error (please refer to Technical Annex, page 33). Instead, Milestone M2.1.2 is intended.

3. Subtask WP2.T1.4: The text:

> "By comparing our findings to those in the literature related to *temporal lobe modelling*, we expect to gain more insight into adequate means to represent auditory stimuli and memory (M2.1.4)."

should read:

> "By comparing our findings to those in the literature related to *the temporal aspects of cognitive speech processing*, we expect to gain more insight into adequate means to represent auditory stimuli and memory (M2.1.4)."

Task 2:
Discovering patterns in the output of WP1 has not been attempted yet since work has concentrated on understanding CMM, and the use of the CSSR algorithm. CMM theory and tools to enable incremental learning have not yet been investigated either since a debugged and reliable implementation of CSSR is first required. This work is currently being pursued by WP2. Once an in-house implementation of CSSR is available, and its sensitivity to noise addressed, outputs from WP1 (as well as WP2 segment clustering, and

other WPs, if found advantageous) will be fed to the modified CSSR algorithm. Milestone M2.2.1 has therefore been moved forward to M15. Other approaches similar in nature to CMM are also being explored.

Finally, a typographical error that concerns the entire WP2 deliverable can be found on page 52 of the Technical Annex. In the "Deliverables list" table it can be seen that the delivery date for D2.1 (this deliverable) is stated to be due at M24. This should read M12, i.e., at the end of the first year of work.

4. List of deliverables

Table 3.2.1: Deliverables List for WP2

Del. no.	Deliverable name	Workpackage no.	Date due	Actual/Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months *)	Lead contractor
D2.1	PD/DME Modules. Applicability of CMM Learning for ACORNS	WP2	M12	M12			TKK

***) if available**

Table 3.2.2: Milestones List for WP2

Milestone no.	Milestone name	Workpackage no.	Date due	Actual/Forecast delivery date	Lead contractor
M2.1.1A	Coarse PD	WP2	M6	M12	TKK
M2.2.1	Applicability of CMM Learning for ACORNS	WP2	M9	M7, M10, further results will be reported in M15	TKK
M2.1.1B	Auditory pre-processing with DMEs	WP2	M12	M15	TKK
M2.1.3	Temporal Structures	WP2	M24	M12	TKK
M2.1.5	Self Directed Search	WP2	M30	M12 (partially)	TKK

3.4 WP3 Memory Organisation and Access

As part of WP3 of the ACORNS project there has been the design of a first stage memory architecture to fulfil the requirements for the other work packages in the project. This incorporated an examination of memory models that have been developed in the past for various associated research fields such as biology, psychology, neuroscience, engineering and computation. From this study a preliminary memory model was outlined that has the capacity to perform pattern storage, discovery and retrieval within an overall architecture based on the memory-prediction model of Hawkins. In addition to the examination of the prediction-memory model the report looks at those memory capabilities found by neuroscience studies that have proved fundamental for performing of higher cognitive tasks such as speech recognition for inclusion in the ACORNS memory architecture. This work package report concentrates on the biological basis of cognitive memory functions such as attention, working memory, episodic memory, sensor/motor grounding systems, reinforcement learning as well as previous computational memory/learning models of such memory

functions. There was also consideration of how they might be combined into the memory architecture for the project to develop the preliminary memory architecture. The preliminary memory model developed for the ACORNS projects incorporates the different features of memory models considered in the report. For instance the episodic memory binds multimodal inputs to achieve an associator network that is able to take a memory fragment (a section of the input) and from this recreate the full memory. In addition to the production of this report in WP3, progress has been made on the implementation of a computational attention/working memory model. This model makes use of the auditory features produced in WP1 and WP2 to produce a bottom-up saliency map in order to represent speech attention based on lower level features.

Table 3.3.1: Deliverables List for WP3

.

Del. no.	Deliverable name	Workpackage no.	Date due	Actual/Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months *)	Lead contractor
D3.1	Report focussing on the memory architecture requirements	3	30/11/07	28/11/07	12	6	USFD

***) if available**

- List of milestones, including due date and actual/foreseen achievement date

Table 3.3.2: Milestones List for WP3

Milestone no.	Milestone name	Workpackage no.	Date due	Actual/Forecast delivery date	Lead contractor
3.1.1	Memory architecture requirements defined	3	31/5/07	31/05/07	USFD
3.1.2	Description of the memory architecture software implementation available	3	30/11/07	30/1/08	USFD

Difference from schedule for milestone 3.1.2

Although progress has been made on the implementation of the memory architecture, the software is not yet complete. However, work on a model of attention and working memory has been started. Given that the post-doc at the Sheffield partner was not recruited until 7.5 months after the project was started this implementation is slightly behind. However, we do not anticipate any problems with catching up this lost time and the next milestone for this workpackage being on time in 31 May 2008.

3.5 WP 4 Information discovery and integration

1. Workpackage objectives from the TA:
    1. To develop information discovery and integration mechanisms.
    2. To study how content addressable memory can be used for information representation and access.
    3. To investigate how to associate speech features and patterns with speech events and evidences.
    4. To integrate exemplar-based matching and high-dimension salient feature representation for access.

We believe these objectives are still relevant.
The staring point for this period is expertise of the partners in this field as well as the ideas that were developed and described in the TA.

2. Progress towards objectives
We focused on the further development of the idea of positivity constraints as mentioned in the TA in the context of LSA/SVD-based (LSA=Latent Semantic Analysis; SVD=Singular Value Decomposition) dimension reduction. This lead to an investigation in the application of NMF-based learning (NMF=non-negative matrix factorization). Initial experiments used phone recognition output as a substitute for the output of WP2 in an attempt to find recurring patterns (words) in speech. These experiments (see Stouten, Demuynck, Van hamme, 2007) turned out to be so successful that we attempted to integrate phone discovery and word discovery in one unsupervised learning experiment. The result is described in Van hamme, 2007. Using the ACORNS Dutch database, this experiment shows that:

- The task of discovering a vocabulary of at least 10 words (13 actually) was achieved (1st objective)
- A successful method to associate speech features and evidences in other modalities is given (3rd objective)
- High-dimensional feature representations can be used for recognition/access (4th objective). However, at this moment, we still need to investigate to which extent the internal representations can be thought of as a collection of exemplars or rather models with a large number of parameters

Similar results were obtained on the ACORNS Finnish database.
Meanwhile, we also worked on information and pattern discovery algorithms based on the multigrams approach of Deligne and Bimbot. In this framework, symbolic input strings are explained as a sequence of multigrams, which have a non-unique symbolic realization (much like an HMM has a hidden state emitting non-unique symbols). This work started out as a baseline to compare the NMF-based results with. But the approach needed extensions in order to make the comparison fair (o.a. the use of lattice inputs instead of single best results). At this moment, it is a viable unsupervised learning technique for achieving the ACORNS goals.

3. Deviations from the work plan:
After the successful investigations for unsupervised information discovery and given the required input from WP3, we decided to swap task 1 or WP4.1 (content-addressable memories) and task 2 or WP4.2 (LSA-representations of speech events). This change was discussed on June 28, 2007 with the project officer.
This change also has an impact on M4.1.1 through M4.1.3 (activation/verification mechanisms), which are the milestones related to task 1. However, given the models developed in task 2, it is already clear that the activation/verification mechanisms need not necessarily be structured in the 3 given layers: acoustic, word and semantic. The texts in D4.2 clearly show that the interconnection of these levels mentioned in the TA is even stronger than assumed. At this point, D4.2 shows how word-level (or rather tag-level) activations are computed (which most closely matches M4.1.2) and can be used for verification by simple thresholding. But D4.2 shows that there is no need to form an explicit acoustic (level 1) activation/verification mechanism. Moreover, models at the highest level of abstraction (tags in the current ACORNS database) integrate evidence of the acoustic and word levels without clearly segmenting at the lower levels. During the writing of the TA, we assumed that segmentation at all lower levels is a necessary condition for speech analysis or recognition, but D4.2 proves this is not the case. We therefore propose to recast task 1 and its associated milestones as outlined below.

WP4 - Task 1.
Content Addressable Memories map some “key” into a “value”. In the context of the ACORNS project, this “key” will be an ensemble of events or co-occurrence of events happening in a space of low abstraction. The keys are events in a space with a higher level of abstraction. Both the input and output space of this mapping are not necessarily discrete with binary activation (like the absence or presence of a symbol as used in e.g. CAMs for network routing), but are fuzzy in the sense that multiple events may be activated simultaneously and to a variable extent. In task 1, we will build a hierarchy of such fuzzy mappings such that the composite mapping finally maps the acoustic space to a semantic space. The number of levels in the hierarchy needs to be determined as well as the dimension of the feature space of the intermediate layers. Like in a mammal brain, mappings may take input from several layers and upstream as well as downstream processing may be required. The actual implementation of the mapping depends much on the dimension of the input and output space, distance metric and its description (parametric or exemplar-based).

M4.1.1 – M12. Implementation of activation mechanisms at the tag level.
In this milestone, we want to show an integrated mapping from acoustic events to "semantic" tags as defined in the "year 1" ACORNS databases. Since the mapping is fuzzy, acoustic input leads to an activation of events in the output space (the tags).

M4.1.2 – M24. Top-down learning of patterns and computation of activations.
Vocabulary acquisition in humans is a top-down process in that the global structure is learned first and the finer structure emerges later. Here we develop a learning paradigm where the finer structure (e.g. phone-like units) emerge from the larger units (e.g. words). Hence, the lower level of abstraction (acoustics, phones) is organized after learning at the higher levels (words, semantics) has progressed to acceptable levels. We will develop a method for segregating layers of representation into organized sublayers, in which *activations* of newly emerged events can be computed.

M4.1.3 – M36. Activation/verification mechanisms in hierarchical CAMs.
In this phase, we will optimize the implementation of the mappings (CAM) between the representation layers.
Within this model, *activations* will be computed from events at the lower levels, which will be *verified* by the models at higher levels and may be used for feedback to the lower levels. Some options for this implementation are: (1) the roadmap algorithm, which is appropriate for a wide variety of distance metrics and exemplar-based learning, even in high dimensional spaces, (2) decompositions into a dictionary of additive parts, which incorporates some "lateral inhibition", (3) weighting of the input space as outlined in the TA and (4) neural networks, … The work for this milestone will be integrated in D4.3.

4. List of deliverables

Table 3.4.1: Deliverables List for WP4

Del. no.	Deliverable name	Workpackage no.	Date due	Actual/Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months *)	Lead contractor
D4.1	Implementation and test of activation-verification mechanisms	WP 4	M12 = 30 Nov 2007	M24 = 30 Nov 2008			KUL
D4.2	Report on LSA representation and SVD dimension reduction	WP 4	M24 = 30 Nov 2008	M12 = 30 Nov 2007			KUL
D4.3	Report on exemplar-based and activation-based matching	WP 4	M36 = 30 Nov 2009	M36 = 30 Nov 2009			KUL

***) if available**

Table 3.4.2: Milestones List for WP4

Milestone no.	Milestone name	Workpackage no.	Date due	Actual/Forecast delivery date	Lead contractor
M4.1.1	Old description: Activation-verification mechanism implemented on 1st layer. New description: Implementation of activation mechanisms at the tag level.	WP 4	M12 = 30 Nov 2007	M12 = 30 Nov 2007	KUL
M4.1.2	Old description: Activation-verification mechanism implemented on 2nd layer. New description: Top-down learning of patterns and computation of activations	WP 4	M24 = 30 Nov 2008	M24 = 30 Nov 2008	KUL
M4.1.3	Old description: Activation-verification mechanism implemented on 3rd layer. New description: Activation/verification mechanisms in hierarchical CAMs	WP 4	M36 = 30 Nov 2009	M36 = 30 Nov 2009	KUL
M4.2.1	LSA representation and SVD dimension reduction	WP 4	M24 = 30 Nov 2008	M12 = 30 Nov 2007	KUL
M4.2.2	ASMs defined from WP1 and WP2 features and automatic segmentation	WP 4	M36 = 30 Nov 2009	M36 = 30 Nov 2009	KUL
M4.3.1	Time synchronous exemplar-based and activation based matching	WP 4	M24 = 30 Nov 2008	M24 = 30 Nov 2008	KUL
M4.3.2	Time-asynchronous matching and non-Euclidean distance	WP 4	M36 = 30 Nov 2009	M36 = 30 Nov 2009	KUL

3.6 WP5 Interaction and communication

The research in this WP is structured in four tasks.

Task 5.1 Creation of a platform for learning in the memory-prediction framework
In this task we will create the basic software environment that is needed to integrate the modules produced in WP1 – WP4 and to conduct experiments with language learning. The RUN has provided the part of the

system that generates the agent's responses. The platform will come in two versions: one for off-line experiments, and one that can be used for demonstrations.

Task 5.2 Multimodal integration
This task is dedicated to the development of procedures and software for the integration of speech input and visual input for disambiguating spoken utterances and feedback that is equivalent to hugging.

Task 5.3 Architecture for interaction
In this task we will design and implement a fully operational system that can conduct a multimodal dialogue, using perception-action loops on several parallel levels. Loops at the lowest level cater for latency-free communicative responses, without the need for parsing the semantic contents of an utterance. On the highest level the system must be capable of conscious reasoning.

Task 5.4 Experiments with language learning
Three major experiments will be performed, corresponding to three stages of language learning. In the first stage the system will learn basic communicative behaviour, mainly to show that is can engage in interaction. In the second phase the system will acquire a basic vocabulary, resulting in the emergence of sub-word units. In the third experiment the system will learn a larger vocabulary and basic rules of syntax.

WP5 objectives as specified in the TA – focus on first year

In this WP we integrate the processing, representations and technologies developed in the other WPs, and we will add purposeful behaviour aimed at learning to communicate and multimodal integration, which is found to enhance word segmentation significantly (Roy and Pentland, 2002).

Task 5.1: Purposeful learning to communicate
We assume that our learning agent (Little Acorns) is endowed with an innate urge to communicate with people in her environment, especially her caretakers. We simulate this urge by designing Little Acorns such that she will attempt to maximise the value of a function equivalent to 'caretaker attention' or 'caretaker appreciation'. This function to be optimised becomes more complex (multi-dimensional) in later stages, as learned behaviour becomes more complex.
In the first year of the project we will design a target function of a number of observable variables, such as the frequency and intensity of positive and negative feedback on the behaviour of Little Acorns that is elicited by the combined auditory and 'visual' input. The target function must have a sufficient degree of human credibility, i.e., the criteria against which it is maximized must reflect what happens in the life of babies.
Furthermore we will extend existing learning mechanism to find the best possible interpretations of the input, generate a response, and update the memory contents according to the feedback from the environment. The agent's responses to the input will be cast in symbolic terms, specific enough to express intentions. We assume that for the implementation of a successful learning procedure it is not necessary to enact the intentions in detailed physical gestures.
At the end of the first year the purposeful learning behaviour will be demonstrated by showing that the agent can learn to distinguish between a number of relevant, yet simple acoustic/visual events.

Task 5.2: Multimodal Integration
Realistic communication is of pivotal importance. Therefore, we will design, build and test procedures that enable Little Acorns to integrate inputs from several parallel channels. In order to do this, separate representations will be formed of the inputs in the individual channels. In WP5, the focus will be on developing cognitively plausible representations of the inputs in the other channels. Methods will be developed for creating associations between representations in parallel channels (such as vision and hugging), and for using conditional probabilities of co-occurrence of specific patterns in parallel channels to enhance learning.
During the first year of the project research will focus on developing methods for representing visual inputs in a probabilistic manner that is compatible with the basic tenets of the memory-prediction theory.

In addition, mechanisms must be designed that enable the creation of dynamic links between representations in the visual memory and the auditory memory. For this purpose we will encode visual input in symbolic representations, so as to avoid the need for developing full-fledged processing of input signals in other modalities than audio. However, the way in which memory representations of the visual features of objects and actions are formed must be compatible with the general assumptions underlying the memory-prediction framework.

During the first year of the project we will limit ourselves to forming representations of static objects and events that are easy to locate in time and space (e.g., footsteps on the stairs). At the end of the first year we will demonstrate that such representations can be formed in a plausible manner, and that representations can be linked to form multimodal concepts.

Task 5.3: Architecture

To create an environment in which an artificial agent can learn to communicate we will develop a multi-layered dialogue system somewhat similar to Thórisson (2002). The system's architecture is in accordance with the assumption that learning results in a structure of layered perception-action loops. Our agent will be able to produce behaviour meant to attract the attention of its environment and show that it has noticed that it is being addressed by somebody. This will be accomplished by perception-action loops on a low level in the cortical hierarchy, where links can be established that do not require semantic interpretation of the input. On successively higher layers perception-action loops will be learned that do require linguistic interpretation of the input. Yet, repetitive and familiar situations will be handled 'from memory', i.e., without an explicit reasoning stage during which conscious decisions are being computed. Such automated reactions will emerge as the result of repeated association between a given input pattern and an ensuing action. At the highest level of the hierarchy explicit reasoning will be applied to compute appropriate reactions to unfamiliar inputs. One example of unfamiliar input is the occurrence of new words, which must be linked to new concepts.

The architecture of an artificial agent that is capable of learning adequate auditory communication behaviour is radically different from the conventional automatic pattern recognisers and dialogue systems.

The system that we envisage will be based on the memory-prediction theory, which assumes plastic and evolving links between patterns in the sensory input and responses.

The architecture will also contain a module that models the evolving needs of the agent. This module will drive the dialogue engine, in such a manner that it can also guide the learning process. As more objects are introduced in the agent's world, there will be a need to learn new words and new sub-word units.

Task 3 is primarily devoted to the software engineering aspects of the integrated learning agent. The most important objective of the task in the first year of the project is to define and implement interfaces between all modules and software packages that must yield input to or produce output of the learning agent. At the end of the first year this must result in a stable software package that will be implemented by all partners in the consortium, so that they can run experiments and demonstrate the system.

Task 5.4: Experiments

We will conduct three experiments, simulating three subsequent stages of language acquisition. The first year only concerns the first stage in this learning process.

In the first stage of her development Little Acorns must learn to understand that she is being addressed, and that her caretaker expects some response that shows that his presence is being recognised. To that aim we will train Little Acorns to learn to recognise her name and to learn to understand about ten simple nouns, such as 'papa' and 'mama', 'eat', 'drink', and a couple of words related to washing and changing diapers. Little Acorns must to distinguish between speech addressed to her and speech that may be going on in the background. In order to enable her to associate child-addressed speech from other noises, the speech addressed to Little Acorns must be accompanied by some kind of additional input. The exact nature of that additional input is probably not very important, as long as it can be cast in such a manner that a couple of different domains (food, washing, caressing) can be distinguished. It is, however, necessary to make the additional input somewhat noisy, so as to simulate realistic learning environments.

For this experiment we will record training speech from four speakers in Dutch, English, Swedish and Finnish. The utterances will refer to the same extra-linguistic concepts in all four languages. The utterances to be recorded will be modelled after the type of speech that caretakers use to address babies during the first

months of their lives, up to the stage where normally developing babies start babbling. For each word that the system has to learn we will record some 100 utterances from each of the four speakers that will be used to train the system. In addition, we will record some ten tokens of the utterances from all speakers for testing the system. Care will be taken to obtain a realistic degree of variation in syntax, speaking rate and intonation in all utterances. All speech for training and testing the system will be recorded in quiet environments. Realistic background noise can then be added later on. In addition to child-addressed speech we will also obtain recordings of 'background speech', mainly from speakers other than the four 'caretakers' in each language. For each of the concepts expressed by the utterances we will develop a set of corresponding representations of visual representations. Experiments will compare learning behaviour and the resulting words (and perhaps sub-word units) that result from training with different voices in individual languages as well as between languages.
The costs of making and annotating the recordings will be so low that they can be covered by the running overhead costs. The speakers will be volunteers, because the recording protocol does not require special skills. The annotation is not expensive either, because we do not need accurate verbatim transcriptions. It is sufficient to have reliable indications of the 'topic' addressed in the individual utterances.
To enable Little Acorns to learn, utterances from the training database will be played to the system.
Initially, the system must make do with a small number of 'innate' responses, mainly related to showing that it is paying attention. After having heard the same utterance repeatedly, Little Acorns must form representations of that utterance (as well as of parts thereof) in its long-term memory. These representations will link acoustic and visual inputs and will then form responses that will also be stored in long-term memory. This will increase the response repertoire that the system has available. During training the responses of the system will be monitored by the 'caretaker' who will give feedback about the appropriateness of the response. In the first year feedback will consist of a score on a bipolar scale. The scores will correlate with the appropriateness of Little Acorns' responses.
During the first half of the first year the focus will be on the specification of the experiments and the recording of suitable databases for training and testing the system. Intensive interaction with the other work packages is needed during this phase to determine the options and limitations for the experiments, given the state of development of the models that implement input, representation and interpretation of the 'physical' environment. At the same time, the requirements set by the experiments will guide the research in the other WPs. During the last months of the first year the actual experiments will be performed, and the data will be analysed and interpreted.

Progress towards objectives

The activities in WP5 during the first year span all tasks. There have been no swaps in this work package, and all subtasks (5.1-5.4) are addressed in parallel.

For **Task 5.1 Creation of a platform for learning in the memory-prediction framework,** we created the basic software environment that is needed to integrate the modules produced in the other work packages. This was done in collaboration with KUL. The RUN has provided the part of the system that generates the carer's responses and the module that takes care of the interaction between the carer and the learner. The platform is based on MATLAB and comes in two versions: one demonstrating the learning capabilities of the learner in batch mode, the other demonstrating the capabilities of learner during incremental training. The current version of the carer-learner interaction is now able to show the learner able to learn 13 words based on multimodal stimuli.
The interaction between carer and learner is based on a broadcasting messaging protocol involving call-back time in all dialogue events, which allows a flexible use of multi-level responses (meeting task 5.3).

For **Task 5.2 Multimodal integration,** we designed the carer to provide abstract visual tags in parallel to the audio information. The association between the information in the audio and 'visual' domain is hypothesized in the learner, and used to formulate a response by the learner to the carer. Interestingly, the current learning approach shows that this integration can be achieved without explicitly implementing it – instead, it emerges from the information in the stimuli itself. The experiments carried out so far (task 5.4) clearly show that the mapping between audio form, reference and concept can be addressed by the current approach if this mapping is 1-1. (The next stage will investigate more complex mappings).

For **Task 5.3 Architecture for interaction** we designed, implemented and tested a fully operational system that can conduct a multimodal dialogue. Until now, it is based on abstract perception-action loops on one level. This is a direct result of the design of the multi-lingual databases that were recorded to perform experiments (meeting task 5.4). The design of the system itself allows the use of perception-action loops on more than one level, thereby modeling a dialogue system as proposed by Thórisson (2001). This multi-level dialogue model is made possible by implementing the dialogue itself on the basis of a flexible broadcasting messaging protocol involving call-back time in all dialogue events. This task is carried out in close collaboration with the KUL.

For **Task 5.4 Experiments with language learning** we carried out a number of experiments that show how and to what extent the learner is capable of acquiring, storing and reusing representations of words and word-like entities. The experiments reveal interesting aspects of the learning algorithm, but more importantly, give the opportunity to relate model parameters and learning parameters on the one hand to cognitive aspects of human learning on the other hand. For example, the use of memory, the extent to which observed utterances remain accessible in memory, the amount of multimodal data used, the number of training tokens per word, the ordering of the multimodal presentations, speaker dependency, the sensitivity to speech style are all results from the computational simulations which are directly linked to human performance. The document associated with the software D5.1 provides an overview of the experimental details and results.

Deviations from work plan

None. The original plan to hire a PhD for the lifetime of the project has been abandoned for hiring a postdoc for 18 months, which will boost the research on building semantically oriented representations and strengthen the cognitive plausibility and interpretation of the experiments. This will be effective M12-M30.

List of deliverables and milestones

Table 3.5.1: Deliverables List for WP5

Del. no.	Deliverable name	Workpackage no.	Date due	Actual/Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months *)	Lead contractor
D5.1	System demonstrating the capacity for acquiring language and communication skills	5 (1, 2, 3, 4)	12	M8 version 0, later updates at M9, M10, M11	7 RUN 2 KUL		RUN
D5.1	Report on System demonstrating the capacity for acquiring language and communication skills	5 (plus 4)	12	Version 2	1		RUN

Table 3.5.2: Milestones List

Table 3.5.2: Milestones List Milestone no.	Milestone name	Workpackage no.	Date due	Actual/Forecast delivery date	Lead contractor
M5.1	Specification of first year experiments	5	3	3 (5, 6 follow up versions)	RUN
M5.2	Specification of cost function for well being	5	6	6 (7, 9, 11 follow up versions)	RUN
M5.3	Complete implementation of basic learning system	5	10	8 (9,10,11 follow up versions)	RUN

3.7 WP 6 dissemination and Use

There are five tasks in this workpackage.

The first task is related to the establishment of a public website. This web site, www.acorns-project.org, was operational in due time.

The second task relates to organising a workshop dedicated to topics of ACORNS. Although the first workshop was planned for Month 18, we had the unique opportunity to organise a high level workshop in 2007, because additional funding was provided by the European Science Foundation (ESF) for an exploratory workshop, and by the Dutch National Science Funding (NWO) for international activities. We were able to invite 25 world leading scientists in the fields of language acquisition, speech recognition, language evolution, and psycholinguistics. Among these 25 scientists were also seven of the eight SAC members of ACORNS: Prof. Cutler, Prof. Fikkert, Prof. Svendsen, Prof. Daelemans, Prof. van Compernolle, Prof. Lee, Dr. Rougier. Unfortunately, J. Hawkins was not able to accept the invitation. At the workshop, also the workpackage leaders of ACORNS were present and the first year results of ACORNS were presented. Feedback was received by the SAC members, as well as the other participants.

The third task relates to open source software. This task is planned for the final year.

The fourth task deals with the publications in ACORNS. Given the fact that only fundamental scientific research is done in this project, this tasks mainly relates to encourage writing of published papers. Four papers were published during the first year of ACORNS and more papers are planned for year two based on the results of the first year.

Task five is devoted to spreading awareness beyond the scientific community. In the Netherlands a press release was made for ACORNS. This resulted in one page article in the periodical of the Radboud University, VOX (number 18, 31th May 2007). For the next two years we plan to write updated press releases, and we expect more reactions by the press, because it will be easier to show clear results.

Table 3.6.1: Deliverables List

Del. no.	Deliverable name	Workpackage no.	Date due	Actual/Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months *)	Lead contractor
D6.1	Public Website operational	6	M3	M3	1	1	RUN
D6.2.1	First Project Workshop	6	M18	M12	1	1	RUN
D6.4	Published papers	6	M12	M12			RUN
D6.5	Public Awareness	6	M36	M6			RUN

***) if available**

- List of milestones, including due date and actual/foreseen achievement date

For the milestones, see the deliverables.

4 Consortium Management

- **Consortium management tasks**

The management tasks for the first year went relatively smoothly. The meetings were organised as planned, the minutes of the meetings and audio conferences were always sent in time. The only difficult issue was getting the Consortium Agreement in place. This Agreement was only signed in March 2007. It took some mails and discussions with the IPR representatives of especially TKK and K.U. Leuven to settle the important issues on property rights and pre-existing know-how.

- **Contractors**

It took some time to fulfil all positions in the project. In July 2007 almost all PhD and postdoc positions were fulfilled. This fact did not result in a slow start for the project as a whole, since all partners found temporal solutions to get the work done. Generally speaking, the consortium is very dedicated to the project, the PhD's and postdocs are very enthusiastic and motivated and also the senior staff members spend relatively much time to the project, because they like the topics and challenges. The project meetings were very useful and constructive meetings and always clear appointments were made for the next period.

No changes in responsibilities were necessary.

For workpackage 4, it turned out to be more reasonable to start in the first year with Task 4.2 (LSA representations of Speech events) and to deal with Task 4.1 (Content Addressable Memories) in the second year of the project. The PCC agreed with this relatively minor change in the workprogramme.

- **Project timetable and status**

WP		Month / Task	1	2	3	4	5	6	7	8	9	10	11	12
WP0	Project Management	T0.1	M			M			M				M	M/D0.2
		T0.2						D0.1						D0.3/D0.4
WP1	Signal Represent-ations	T1.1									M1.1			M1.2 D1.1
		T1.2												
WP2	Signal Patterning	T2.1												M2.1.1A/D2.1
		T2.2										M2.2.1		
WP3	Memory Organization and Access	T3.1						M3.1.1						D3.1
		T3.2												
		T3.3												
		T3.4												
		T3.5												
WP4	Information Discovery & Integration	T4.1												M4.1.1
		T4.2												M4.2.1/D4.2
		T4.3												
WP5	Integration and Communic-ation	T5.1												
		T5.2										M5.3		
		T5.3						M5.2						
		T5.4			M5.1									D5.1
WP6	Dissemination and Standardization	T6.1			D6.1									
		T6.2												D6.2.1
		T6.3												
		T6.4												D6.4
		T6.5						D6.5						

Updated Gantt Chart for Year 1

- **Co-ordination activities**

The Acorns project started on the first of December 2006. Five one and a half day lasting project meetings took place during the first year:

1. Kick-off meeting in Nijmegen 20th and 21st of December 2006
2. Second project meeting in Sheffield 19th and 20th of March 2007
3. Third project meeting in Fiskars 13th and 14th of June 2007
4. Fourth project meeting in Stockholm 4th and 5th of October 2007
5. Fifth project meeting in Leuven 29th and 30th of November 2007

Audio conferences were scheduled to prepare the third, fourth, and fifth meeting:
1. 24th of May 2007
2. 5th of September 2007
3. 16th of November 2007

In between the 'official' meetings and conference calls, a very lively exchange of e-mail traffic and data took place. ACORNS is a relatively small project which makes it relatively easy to manage. Prof. Unto Laine invited the project members to Fiskars (June) , a small, nice village in the neighbourhood of Helsinki. This meeting was very favourable for team building in ACORNS, resulting in frequent informal exchanges of information and data after this meeting and beyond.

On 3 October 2007 a meeting was held at KTH to discuss issues in Computational Mechanics Modelling, with representatives from KTT, KTH and RU Nijmegen.

Several meetings (mostly for part of a day) have been held n which RU Nijmegen and KU Leuven discussed issues related to the platform for conducting experiments.

5 Annex A: Plan for dissemination and Use

5.1 Exploitable knowledge and its Use

Table 5.1.1 Overview table of exploitable knowledge

Exploitable Knowledge (description)	**Exploitable product(s) or measure(s)**	**Sector(s)** of application	Timetable for commercial use	Patents or other IPR protection	**Owner** & Other Partner(s) involved
1. Procedure for blind bottom-up speech segmentation		1. Speech recognition 2. Industrial inspection; signature analysis	2010	patent application (FIN-20075696) filed	TKK
2. Software package for speech signal processing		1. speech recognition and speech coding	After 2010		KTH
3. Structure detection by means of Non-Negative Matrix Factorisation		1. Speech recognition 2. Data mining	After 2010		KU Leuven

1. Procedure for blind bottom-up speech segmentation using Discrete Model Elements

- Discrete Model Elements (DME) are a novel approach for finding local structure in continuously changing signals. Examples of such signals are speech, but also noise and vibration signals produced by machinery, natural systems, etc. The goal of bottom-up segmentation is to find points in time where the behaviour of the system generating the signals changes significantly, suggesting that the system is making a transition from one state to another.
- Exploitation of the segmentation procedure will be pursued mainly by the originator, i.e., TKK. The other partners will assist TKK in contacting commercial companies.
- Commercial exploitation will probably depend on finding commercial companies interested in developing the basic results obtained so fat into an operational software module.
- Actual deployment of the novel procedure will require additional research, among others to better understand the robustness of the procedure against additive and convolutional noise.
- TKK, the originator of the novel procedure, has filed a patent application (FIN-20075696)

2. Software package for speech signal processing

- The package contains tested software modules for conventional signal processing, primarily for use in the consortium, to guarantee that there are no differences between the results of processing identical input by different partners. The procedures can also be used outside the consortium.
- The procedures have been implemented by KTH; the other partners have assisted KTH by rigorously testing the code. For the moment no commercial applications are foreseen.
    - We see the major application of the software in scientific research, where there is a need for tested and verified procedures for basic speech signal processing routines.

3. Structure detection by means of Non-Negative Matrix Factorisation

- Non-negative Matrix Factorisation (NMF) is a novel technique for discovering structure in matrices describing observations form physical processes, represented in terms of non-negative numbers (eg. Energies, number of occurrences, etc.).We have developed NMF to detect structure in continuous speech, based on a representation that tracks the number of transitions between labels after vector quantisation.
- The work has been carried out mainly by KU Leuven.
    - For the time being, we expect that the knowledge will mainly be used in the ACORNS project.
- Further additional research and development work, including need for further collaboration and who they may be;

5.2 Dissemination of knowledge

Table 5.2.1 Overview table of dissemination activities

Planned /actual Dates	Type	Type of audience	Countries addressed	Size of audience	Partner responsible /involved
15/01/2007	Press release	General public	Netherlands	16 Million	RU Nijmegen
	Media briefing	*Higher education*			
26/11 – 28/11/2007	Workshop	Research	Europe	35	RU Nijmegen and USFD
	Exhibition	*Industry (sector x)*			
Several dates	Publications; for details, see below	Scientific	global	15,000	all
01/02/2007	Project web-site	General Public, but mainly scientists	global	millions	RU Nijmegen
	Film/video				

A Press Release describing the project was issued in the Netherlands on 15 January. This resulted in a limited amount of media coverage.
Additional press releases are planned in the future, when demonstrable results are available.

The project website (http://www.acorns-project.org) has been released in the middle of February 2007. The website is being kept up-to-date by the project coordinator.

An on-invitation-only workshop has been conducted in Leuven, supported by the European Science Foundation and the Dutch science organisation NWO. A short report on this workshop will follow.

List of publications:

Regular Papers:

Lou Boves, Louis ten Bosch, Roger Moore "ACORNS -- towards computational modeling of communication and recognition skills", Proc. ICCI-2007.
Veronique Stouten, Kris Demuynck, Hugo Van hamme "Automatically Learning the Units of Speech by Non-negative Matrix Factorisation",
Proc. Interspeech 2007.
Veronique Stouten, Kris Demuynck, Hugo Van hamme "Discovering Phone Patterns in Spoken Utterances by Non-Negative Matrix Factorization", IEEE Signal Processing Letters 2008

Louis ten Bosch, Bert Cranen "A computational model for unsupervised word discovery", Proc. Interspeech 2007.
Hugo Van hamme "Non-negative Matrix Factorization for Word Acquisition from Multimodal Information Including Speech", ESF Workshop, Leuven November 2007.

Theses:

Okko Räsänen "Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture", MSc Thesis, Helsinki University of Technology, Espoo, November 5, 2007.
Alexander Bertrand "Zelflerende Spraakherkenning via Matrix-factorisatie", Katholieke Universiteit Leuven - Departement Elektrotechniek ESAT, 2007, [in Dutch].

5.3 Publishable results

At the end of the first year there were no software modules that are sufficiently tested and documented for public release.