



Project no. 034362

ACORNS

Acquisition of COmmunication and RecogNition Skills

Instrument: STREP
Thematic Priority: IST/FET

Periodic Activity Report

Period covered: from 1 December 2008 to 30 November 2009

Date of preparation: 1 December 2009

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

Project coordinator name: Prof. Lou Boves
Project coordinator organisation name: Radboud University, Nijmegen
Revision Final

Table of Contents

1. Executive summary	1
1.1 Summary description of project objectives	1
1.2 Work performed	2
1.3 Results achieved	2
How do babies learn language	2
Implications for ASR	3
Software and corpora	3
Publications	4
2 Project objectives and major achievements during the reporting period	1
2.1 General Project Objectives	1
2.1.1 Objectives for the reporting period	1
2.2 Work Performed	2
2.3 Results	2
2.3.1 WP1	3
2.3.2 WP2	3
2.3.3 WP3	3
2.3.4 WP4	4
2.3.5 WP5	4
2.3.6 WP6	4
2.4 Addressing the recommendations from the second year review	5
3 Workpackage progress of the period	6
3.0 Workpackage 0 Project Management	6
3.1 WP1 Signal Representations	7
3.1.1 Workpackage objectives and starting point of work at beginning of reporting period	7
3.1.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved	7
3.1.3 Deviations from the project work programme, and corrective actions taken/suggested	8
3.1.4 List of Deliverables	8
3.1.5 List of milestones	9
3.2 WP2: Signal Patterning	10
3.2.1 Workpackage objectives and starting point of work at beginning of reporting period	10
3.2.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved	10
3.2.3 Deviations from the project work programme, and corrective actions taken/suggested	13
3.2.4 List of Deliverables	13
3.2.5 List of Milestones	14
3.3 WP3 Memory Organisation and Access	16
3.3.1 Workpackage objectives and starting point of work at beginning of reporting period	16
3.3.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved	16
3.3.3 Deviations from the project work programme, and corrective actions taken/suggested	17
3.3.4 List of Deliverable	17
3.3.5 List of Milestones	17
3.4 WP 4 Information discovery and integration	19
3.4.1 Workpackage objectives and starting point of work at beginning of reporting period	19
3.4.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved	19
3.4.3 Deviations from the project work programme, and corrective actions taken/suggested	21
3.4.4 List of Deliverables	22
3.4.5 List of Milestones	22
3.5 WP5 Interaction and communication	23
3.5.1 Workpackage objectives and starting point of work at beginning of reporting period	23
3.5.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved	24
3.5.4 Deviations from the project work programme, and corrective actions taken/suggested	26
3.5.4 List of Deliverables	27
3.5.5 List of Milestones	27
3.6 WP 6 dissemination and Use	28

<u>4 Consortium Management.....</u>	<u>30</u>
<u>Annex 1: Plan for dissemination and Use</u>	<u>1</u>
<u>1 Exploitable knowledge and its Use.....</u>	<u>1</u>
<u>2 Dissemination of knowledge.....</u>	<u>4</u>
<u> List of publications.....</u>	<u>4</u>
<u> Planned presentations and publications.....</u>	<u>7</u>
<u>5.3 Publishable results.....</u>	<u>7</u>



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 (coordinator)
- 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 aims at testing the viability of the memory-prediction theory of intelligent behaviour as a basis for modelling the acquisition of language and communication skills. The input for the model consists of audio signals in combination with representations of the context to which spoken utterances refer. The project work plan is structured in five technical work packages, four of which are devoted to several aspects of sensory processing, discovery of meaningful structures and associations in the input data and the representation of learned structures in memory. The fifth work package is dedicated to integrating results in a comprehensive system and conducting experiments that test the capability of the approach to account for the acquisition of language and communication skills.

The **front-end processing** module, under development in WP1, provides a representation of audio signals that can characterise and process all ecologically relevant sounds and model different sources independently, with a strong focus on features that are important for speech processing.

WP2 focuses on **pattern discovery**, by building computational models that detect recurring patterns in input signals. **Memory organisation and access** is the focus of WP3. We develop computational representations of the different types of memories that are implied in the memory-prediction theory and models of memory and processing resulting from research in psychology. WP4 investigates **Information discovery and integration**, with a focus on emergent representations that result from a combination of bottom-up and top-down signal processing.

For WP5, **Interaction and communication**, the aim is to integrate the results in a system that can simulate the acquisition of language and communication skills. In doing so, emphasis is put on the cognitive and biological plausibility of the procedures, processes and representations developed in the four preceding work packages. Three increasingly more complex stages, one for each year of the project, are defined. At the end of the first stage, our artificial infant had acquired the basic skills needed to understand that it is being addressed; in addition, it had learned some ten words. At the end of the second stage the artificial agent has

ACORNS

learned some 50 words. In the third and last stage of language acquisition that we will simulate, we will investigate how previously acquired knowledge and skills can be harnessed to speed up further learning of language and communication skills. To demonstrate the learning skills we will show that the agent can learn new words when new concepts are introduced in the environment.

1.2 Work performed

Until the start of the third year all experiments were performed with conventional spectral representations in the form of Mel-scale Cepstrum Coefficients (MFCCs). During the third year two sets of new features, based on knowledge about the human auditory system have been developed and tested in several conditions. As a basic proof of concept the novel features were used in 'standard' automatic speech recognition tasks. In addition, the novel acoustic features were used in language acquisition experiments using the learning mechanisms developed in ACORNS.

In addition to novel acoustic features another line of research addressed the way in which the visual features are best represented, and what the implications are of several different ways in which visual input can be made probabilistic, instead of essentially deterministic. It was also investigated how the interaction between the (artificial) learner and supervisor affected learning.

All partners contributed to several sets of experiments aimed at better understanding language acquisition. In the first two years three different approaches to discovering structure in speech signals and learning associations between speech and visual input were developed. In the third year these approaches were further developed, and compared to one another in a learning task that had a second major goal, viz. testing whether learning from multiple speakers helps in understanding new speakers.

Research also continued on the development of memory architectures that are suitable to model language acquisition. Here, the focus was on integrating the major features of the functional memory models developed in psychology and biologically inspired models such as the memory-prediction model.

ACORNS did not only aim at better understanding human language acquisition, but also to use the novel insights for improving automatic speech recognition. In the third year several partners joined effort in investigating ways in which example-based speech recognition could improve upon conventional automatic speech recognition based on Hidden Markov Models.

Finally, the consortium organised a second workshop in the fringe of the international Conference Interspeech that was conducted in Brighton, U.K.

1.3 Results achieved

The research in the ACORNS project has substantially advanced the state-of-the-art in our understanding of human language acquisition. In addition, the work has shown promising directions for improving automatic speech recognition. Last but not least, the project has produced re-usable software tools and data.

How do babies learn language

The research in ACORNS has shown that both the biologically inspired memory-prediction framework and the memory architectures originating from psychological research are heavily underspecified. A major contribution of the project is that we have developed a novel memory architecture that integrates essential aspects from both the biological and psychological approaches. The novel architecture has been implemented and tested for several different learning strategies and mechanisms. Our experiments with this memory architecture have shown that many more behavioural (and brain imaging) data are needed to be able

ACORNS

to evaluate the plausibility of hypotheses and claims about the way in which representations of language units (words, syllables, sounds) emerge as a result of audio-visual communication.

ACORNS has also provided a computational methodology that can be used in future research to address several issues that are being discussed in the literature on human language acquisition. We have been able to show that existing theories of language acquisition lack precision, to a large extent due to the fact that we do not have continuous records of the experience of one or more babies. As a consequence it is not possible to explain newly learned behaviours from experience gained in the time between experiments. We have shown how computational modelling can help to solve this problem.

ACORNS has also shown how computational modelling can shed light on specific issues that are being discussed and investigated in human language acquisition. One example is the question whether learning is more efficient if a baby mainly interacts with only one or two caregivers, or whether learning is enhanced through interaction with multiple caregivers. We have shown that both positions can be maintained, depending on the criteria used to establish the efficiency of early learning and the representations that are being built in memory. For representations that are mainly episodic in nature, interaction with multiple caregivers does not enhance learning of new speakers very much. For representations that are based on co-occurrence statistics of acoustic features interaction with multiple caregivers does facilitate understanding new speakers. The human brain can support both representations. Therefore, it is interesting to investigate the hypothesis that language acquisition involves a combination of multiple different representations.

An important feature of the memory architecture developed in ACORNS is an attention mechanism that modulates the salience of sensory inputs. Learning is enhanced substantially if the learner can select the relevant features in the sensory input. This underlines the need to proceed to experiments in which the (artificial) learner can take a more pro-active role than the learner in the ACORNS project. The ACORNS learner is only partially embodied, in that she can pay selective attention to sensory inputs in a simulated context. In future research the learner should also be given the opportunity to actively invoke sensory input by actively exploring and acting upon the environment. We expect that novel insights in the various ways in which language units can be represented in the brain and will give rise to better diagnose infants that are at risk of developing language and communication disorders, and in the end may result in better tools for remedying these disorders.

Implications for ASR

The ancillary goal of the ACORNS project was to investigate the feasibility of a novel approach to automatic speech recognition. To that end we have investigated whether the internal representations that emerge in a system that learns language as a side effect of learning to communicate differ from the conventional representation of speech as a linear sequence of essentially discrete units, such as sounds, syllables, words, etc.

While it would obviously not be realistic that a three year project could result in an approach that can outperform Hidden Markov Models on which many tens of thousands of person years have been spent for research and development, we can still claim that a novel exemplar-based approach developed in ACORNS holds the promise to remedy some of the intrinsic weaknesses of the HMM approach. Some of the techniques developed for computing the match between previously seen examples and novel inputs (among others distance measures that preserve information about local structure in the acoustic space) might also be carried over to the calculation of observation likelihoods in conventional HMMs.

Software and corpora

ACORNS has produced a set of speech corpora (in Dutch, English, Finnish and Swedish) that can be re-used for the simulation and investigation of procedures for discovering recurrent patterns in speech signals and for creating associations between such recurrent patterns and referents in the physical environment. These corpora will be made available to other researchers.

ACORNS

The ACORNS project has also resulted in a software package that future research can use for investigating the processes underlying the discovery of recurrent patterns that can be associated with meaningful referents. This software will be made available to the research communities in the form of Matlab scripts.

Publications

Up to now, ACORNS has resulted in 3 journal articles, 32 conference papers and 3 master theses, all of which can be accessed through the project's public website www.acorns-project.org

A substantial number of additional journal articles and conference papers is still in the pipeline. These publications will be posted on the public website as soon as they become available. The public website will be updated and maintained for a period of five years after the end of the project.

2 Project objectives and major achievements during the reporting period

2.1 General Project Objectives

ACORNS aims to clarify the feasibility of the memory-prediction theory of intelligent behaviour as it applies to language acquisition and speech communication by developing and testing a computational model of language acquisition informed by the memory-prediction theory. The input for the model consists of audio signals in combination with symbolic representations of the environmental context, to provide grounding. Research in the first year has shown that there is not yet a software implementation of the memory-prediction theory that is powerful enough to build an agent that can simulate a process as complex and as badly understood as language acquisition. More conventional theories of language acquisition and processing suggest that the processes must be modular.

The project work plan is structured in five technical work packages, four of which are devoted to investigating specific aspects of the memory architecture and processing, while WP5 is responsible for the integration of the results and for conducting the experiments that will investigate to what extent the memory-prediction theory can account for the acquisition of language and communication skills. Here we briefly sketch the five work packages:

Front-end processing (WP1) results in a rich internal representation that is suitable to characterise and process essentially all ecologically relevant sounds and to model different sources independently.

Pattern discovery (WP2) investigates computational models that can detect recurring patterns in the input signals and that can be linked to memory representations.

Memory organisation and access (WP3) will focus on 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.

Information discovery and integration (WP4) investigates how patterns in the input signals can be discovered, stored and accessed for the interpretation of novel input signals.

Interaction and communication (WP5) integrates the results of the WPs sketched above in a software platform that allows us to investigate aspects of language acquisition and processing in a setting that resembles human language acquisition: as the result of situated communication between a care giver and a learning agent. Care is taken to ensure that the learning algorithms and memory representations developed in WP1 – WP4 are plausible from a biological and cognitive point of view. For this purpose, it is emphasised that learning should be incremental, and that ‘training’ stimuli are offered to the learning agent only once.

In the first year of the project it has become clear that the borders between the work packages 2, 3 and 4 are fluid and permeable. Therefore, we have drawn up an updated pictorial representation of the structure of the project, which is shown here in Fig. 2.1.

2.1.1 Objectives for the reporting period

Several different algorithmic and architectural approaches to modeling language acquisition have been tried in the first two years. This work resulted in improved insight in many aspects of simulating language acquisition. However, both the Scientific Advisory Committee and the reviewers recommended that the research in the third year should focus on the most promising approaches and to elaborate a detailed qualitative and quantitative comparison among these approaches.

In order to allow a global coherent comparison among approaches, it was decided to define a unique experimental scenario, together with a unique database, and use these as a basis for a formal comparison.

The reviewers also recommended that the consortium should repeat a number of experiments performed in the previous years (that were based on conventional MFCC features) using the novel features under development in WP1.

2.2 Work Performed

Starting from the original Technical Annex, we instantiated the plans for research in the third year in such a manner that we could reap maximal benefit from the recommendations of the Scientific Advisory Committee and the second annual review. In accordance with the conclusion reached at the end of the first year that deeper insight would be obtained from comparing the three approaches developed in the first year (Non-Negative Matrix Factorisation [NMF], Concept Matrices [CM] and DP-ngrams) than from attempting to integrate parts of these approaches in a single model architecture, substantial effort was devoted to developing ways in which the performance of the approaches could be measured and compared. Since it was felt that meaningful comparisons could be performed with the corpus data produced in the first two years, it was decided to refrain from recording new speech data.

In the comparison of the three approaches attention was focused on the trade-off between episodic and more symbolic representations of speech signals in the memory architecture. Closely related to the issue of episodic versus symbolic representations is the degree to which these representations are speaker dependent or –on the contrary- independent of the idiosyncrasies of individual speakers.

In addition to comparing the three approaches to language acquisition we also set out to investigate the impact of the new acoustic features under development in WP1. It was decided to do this by repeating a set of experiments conducted using conventional MFCC features with the new features developed in WP1.

Grounding of the semantics of the speech utterances thanks to their relation to the (simulated) visual inputs was an issue for virtually all work packages. Accordingly, substantial effort has been devoted to the impact of different ways in which the information in the visual channel was represented.

Another issue to address in the third year was the impact of the way in which learner and caregiver interact and especially the frequency and the emphasis with which the caregiver corrects misinterpretations of the learner and the willingness of the learner to trust the input of the caregiver.

A final set of issues to be addressed in the third year relates to the cognitive plausibility of the memory architecture and the processing that takes place on new sensory inputs and previously processed inputs (that are represented in the memory). In this context the scalability of the processing techniques and the memory architecture to larger and more complex tasks (larger vocabulary, more involved syntax, more complex scenes) have also been addressed.

The learning tasks in the third year required the artificial agent to handle a up to 50 words (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.

2.3 Results

In this section we summarise the most important results of the research in the third year. More detailed accounts of the work performed and the results can be found in the deliverables for the third year, all of which have been produced according to schedule.

2.3.1 WP1

We developed a methodology for the design of an effective and computationally inexpensive feature set based on advanced auditory models. In general, we aimed at finding features that retain local distances based on auditory models and results of perceptual experiments. Two novel feature sets were produced, both departing from conventional MFCC coefficients. The first feature set focused on auditory models that can account for distances in static signals, while the second set also accounts for dynamic signal characteristics. In addition, a method was developed and tested to find the optimal subset in conventional MFCCs and the static novel features.

Extensive tests have shown that the novel features outperform conventional MFCC features, especially for speech in adverse conditions. In addition, the novel features have been used in repetitions of structure discovery and learning experiments that were previously conducted with conventional MFCC features. The results of these experiments have shown that the novel features do contribute to the effectiveness of the learning mechanisms.

2.3.2 WP2

Research on discovering structure in speech signals using Concept Matrices (CM), a novel approach based on co-occurrence frequencies associated with semantic concepts, continued. The CM approach was compared with two other learning strategies in a number of experiments. It appeared that the performance of CMs is comparable to the performance obtained with Non-negative Matrix Factorisation (NMF).

A novel method for discovering structure was developed, based on the Permutation Transformation Method and successfully tested on the problem of recognizing intervocalic stop-consonants.

Research on Computational Mechanics Modelling (CMM) for discovering structure in speech signals continued and resulted in a robust algorithm for applying the Causal State Splitting and Reconstruction approach to the type of noisy data characteristic for speech. In addition to algorithmic insights the research also resulted in an open source software package that implements the novel algorithms.

2.3.3 WP3

The most important results associated with WP3 relate to the taking of the ACORNS memory architecture, which was developed over the first two periods of the project, and using it as the framework and focus to inspire the use of approaches not typically associated with speech recognition. This process has allowed us to successfully show the benefits associated with combining various components of the memory architecture (working memory, semantic long-term memory, episodic long-term memory and attention) into speech recognition applications. It has also been possible to show that the memory architecture offers sufficient flexibility to combine diverse approaches within one overall framework. In period 3, existing models such as the attention mechanism used to classify speech/non-speech and the recurrent self-organising map approach for associating speech with semantic feature have been further developed and their performance analysed. Moreover, new models have been developed to assist in the examination and analysis of the ACORNS memory architecture. The first group uses from the memory architecture the interaction between working memory and episodic memory and includes the Temporal Episodic Memory Model (TEMM) for word learning and Acoustic DP-Ngrams for keyword learning. The second group are based on the interaction found in the ACORNS architecture between working memory and semantic memory. These models include a Restricted Boltzmann Machine for spoken word recognition and early vocabulary acquisition using hierarchical NMF. It has been possible to show that WP3 has successfully developed a memory architecture that forms the basis for speech recognition applications and could also be used in alternative research fields. The memory architecture has also allowed us to compare different approaches that are either based on semantic long-term memory and or episodic memory.

2.3.4 WP4

Non-negative Matrix Factorization (NMF) is based on the analysis of co-occurrences of vector quantization (VQ) labels. Therefore, NMF might suffer from information loss that will inevitably occur in vector quantization. Research on several tasks (including the task of recognizing the consonants in noisified VCV-sequences) has shown that increasing the number of codebooks is less effective than applying the soft vector quantization that was already proposed in a 2008 paper.

One of the basic issues addressed in ACORNS is how hierarchical or multi-level representations can be learned. It was shown that NMF can be applied to the output of a previous NMF factorization. Therefore, it appears that the NMF approach is capable of learning multi-level structures.

Another issue that is being investigated in ACORNS is the role of episodic representations (perhaps in combination with more abstract symbolic representations). This question is related to better understanding human speech processing, but also has direct links to improving automatic speech recognition. It was shown that Loopy Belief Propagation combined with a novel Roadmap algorithm for searching matching examples can result in a system that holds the promise of outperforming conventional Hidden Markov Models for continuous speech recognition.

2.3.5 WP5

The research in WP5 concentrated on experiments that addressed the comments and recommendations of the SAC and the reviewers.

First, a series of experiments was conducted to compare the three major learning strategies on a common task, using the same data for learning and comparison. It appeared that the two approaches that represent the acoustic input in the form of co-occurrence statistics of labels produced by a vector quantisation perform differently than DP-ngrams, in which the structure discovery operation is based on the original time-frequency representation. Therefore, it became clear that fundamentally episodic representations act differently in structure discovery than representations that are more symbolic in nature.

Second, experiments were performed to investigate the impact of fuzzy visual/semantic representations on discovering structure and learning. The original crisp visual input was made ambiguous in two different ways: by adding irrelevant visual features and by changing the unique labels (e.g., 'car', 'shoe') into probabilistic scores {car: 0.95, shoe: 0.3, ...}. It was shown that NMF is able to cope with fuzzy visual input, irrespective of the way in which crisp input was made ambiguous.

Third, it was investigated what the impact was of different strategies of the learner in interaction with the caregiver. It appears that learning is not affected substantially by occasional misunderstanding between learner and caregiver, something that may be due to mistakes made by either participant. It was also shown that the learner may override the information about the referents in a spoken utterance suggested by the caregiver if the learner is confident about her own interpretation. These results show that the learning mechanisms can keep operating successfully in less than ideal conditions.

2.3.6 WP6

The major tasks in WP6 were the organisation of the second workshop, the release of the software produced in the project, and the collection of published papers.

The second workshop was organised in the form of a small number of invited presentations, followed by in-depth discussions. It was explicitly decided *not* to organise the workshop in the form of a mini-conference.

To maximise attendance the workshop was planned on the day following the Interspeech Conference in Brighton (11 September). Unfortunately, it appeared that there were many other post-conference activities that competed for attention. Also, several members of the SAC appeared to have obligations in some of these satellite events. Therefore, attendance did not reach the maximum number of 50 participants that we set in advance.

Presentations were given by Lou Boves (on ACORNS), Deb Roy (on related research at the MIT Media Lab), Rochelle Newman (on human language acquisition) and Friedemann Pulvermüller (MRC Cognition

ACORNS

and Brain Sciences Unit, Cambridge, UK) on the neural mechanisms underlying language acquisition. At the end of the workshop a number of potential follow-up activities were identified.

The software for running experiments in interactive learning developed by the ACORNS partners was collected and documented. It is described in more detail in Deliverable 6.3.

In addition to the software the ACORNS consortium will also publish the speech corpora that were collected in the course of the project.

On 30 November 2009, the formal conclusion of the project, ACORNS has resulted in 3 journal papers and 16 conference papers published in year 3. Additional papers are in the pipeline.

2.4 Addressing the recommendations from the second year review

The recommendations from the reviewers and the members of the Scientific Advisory Committee have already been described in section 2.1.1 and addressed in section 2.3.

In addition, the reviewers recommended that we publish in journals and conferences outside the core speech community. As can be seen from the list of publications, several papers were published addressing the computer science, pattern recognition and psychology communities.

3 Workpackage progress of the period

3.0 Workpackage 0 Project Management

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

The tasks for this reporting period were:

1. Manage the project scientifically
2. Prepare, conduct, and report on meetings
3. Update and maintain the internal project website
4. Deliver Periodic Activity Report, Final Activity Report, Dissemination and Use plan, and Management Report
5. Take care of the financial issues in the project
6. Organise Scientific Advisory Board meeting

Achievements:

1. The project was well managed. All major milestones were met, all content deliverables were on time and there has been a good collaboration between the partners. The Activity, and Management reports and Audit certificates, as well as the final versions of the Dissemination and Use plans will be delivered before the end of December 2009.
2. Based on the comments made by the reviewers, we took care that even more integration between the workpackages took place. Integration of WP1 has been accomplished by using the new acoustic features in the three learning environments, and a systematic comparison has been made between the three learning systems. Cross-fertilization and collaboration between the workpackages were stimulated at each project meeting. The results are reflected in the joint publications and the planned joint publications.
3. Four project meetings took place and five conference calls were held.
4. The ACORNS website was updated every few months.
5. The members of the Scientific Advisory Committee were invited to attend the second workshop. However, because of agenda conflicts only one member could attend.

Table 3.0.1: Deliverables List

Del. no.	Deliverable name	WP no.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months *)	Lead contractor
D0.2	Activity Report	0	M36	M37	3	1	RUN
D0.3	Dissemination and Use plan	0	M36	M37	2	1/2	RUN
D0.4	Management Report	0	M36	M37	1	1/2	RUN

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

For the milestones, see the deliverables.

Deliverable D0.3

3.1 WP1 Signal Representations

3.1.1 Workpackage objectives and starting point of work at beginning of reporting period

The main objectives of work package 1 for the year-three reporting period were the completion of deliverable D1.3 and of milestone M1.4.

Deliverable D1.3 consists of the final software modules and a report on the definition of new features and a feature selection method. It is described in the Technical Annex as “*Final Modules for features derived with sensitivity-analysis method criterion, with quantitative evaluation (software and report)*”.

Milestone M1.4 was a midpoint check for progress towards deliverable D1.3. It is described in the Technical Annex as “*New features based on sensitivity analysis method* “. Upon the completion of deliverable D1.2 milestone M1.4 is no longer of significance.

At the start of the reporting period WP1 had completed deliverables relating to Task 2. Furthermore, WP1 had completed the work towards *selection* of features based on the sensitivity matrix approach.

3.1.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved

The objectives for the year-3 reporting period were completed according to the plan described in the ACORNS Annex. Milestone M1.4 was met and deliverable D1.3 was delivered on time. In addition to the planned work, work was performed on *i)* the selection of static and dynamic features, and *ii)* a self-learning algorithm using an NMF-based bottom-up approach to automatically discover, acquire and recognize the words of a language. We discuss the progress in each of the fore-mentioned three topics below.

- In line with the goals for M1.4 and D1.3 we *i)* developed a methodology for the design of an effective and computationally inexpensive feature set based on advanced auditory models, *ii)* developed software modules for these resulting features and *iii)* performed quantitative testing on resulting feature sets. Our methodology is based on the introduction of adjustable parameters in the feature definition. These adjustable parameters are then optimized so that the resulting feature set emulates the behavior of the human auditory system. We considered both static and dynamic aspects of perception. Carefully written software modules were developed. Using these modules, we performed extensive quantitative tests on the resulting feature sets using the standard setup of the well-known HTK recognition toolkit. It was shown that the new features sets provide significant improvement over conventional mel-frequency cepstral coefficients that form the most commonly used feature set for speech recognition. The features were subsequently tested in the context of ACORNS learning algorithms. The results of the latter tests are reported under WP5.
- We extended the work reported under D1.2 for finding a good subset of features for recognition from a larger set using only knowledge of the human auditory system as a measure. In contrast to our earlier work, which considered static measures only, we considered a dynamic spectro-temporal auditory model to select a subset from the standard mel-frequency cepstral coefficients and their velocity and acceleration. We evaluated the selected feature subsets on a real speech recognizer. The results confirm that knowledge of the human auditory system forms a good basis for selecting a subset of features from a larger set of features for the purpose of speech recognition. Our results show that this selection method performs better than established linear-discriminant analysis (LDA) and heteroscedastic linear discriminant analysis (HLDA) methods for dimension reduction. Note that LDA and HLDA can combine input features, and that they use feedback from the recognizer,

ACORNS

whereas our perception base selection cannot combine features, nor can it use classification knowledge. Knowledge obtained from the human auditory system is sufficient to overcome these disadvantages.

- We also introduced a self-learning algorithm using a bottom-up approach to automatically discover, acquire and recognize the words of a language. The magnitude spectrum of a special form of conventional short time Fourier transform (STFT) is used as the input to the algorithm. An unsupervised technique using non-negative matrix factorization (NMF) discovers phone-sized time-frequency patches into which speech can be decomposed. These patches can be considered to be newly developed speech features. Speaker-independent patterns occur in these features and these patterns can also be discovered with NMF. By providing information about the word identity to the learning algorithm, the retrieved patterns can be associated with meaningful objects of the language. In the case of a small vocabulary task, the system is able to learn patterns corresponding to words and subsequently detects the presence of these words in speech utterances.

3.1.3 Deviations from the project work programme, and corrective actions actions taken/suggested

Some additional tasks were performed in addition to the work planned in the project work programme in workpackage 1 for the year-3 reporting period.

3.1.4 List of Deliverables

Table 3.1.1: Deliverables List for WP1

Del. no.	Deliverable name	WP 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	M12	15	10	KTH
D1.2	Modules for a) augmentation of standard spectral features with a stream of milli-second and deci-second features and evaluation on specific phone classification tasks and b) feature selected by sensitivity-analysis method	1	M24	M24	17	13	KTH
D1.3	Final Modules for features derived with sensitivity-analysis method criterion, with quantitative evaluation	1	M36	M36	15	24	KTH

Deliverable D0.3

8

3.1.5 List of milestones

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 millisecond and decisecond features	1	M18	M18	KTH
M1.4	New features based on sensitivity analysis method	1	M30	M30	KTH

3.2 WP2: Signal Patterning

3.2.1 Workpackage objectives and starting point of work at beginning of reporting period

The objectives for WP2's two tasks for the project's third year were as follows:

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

- To complete the one remaining subtask and milestone that was set out for the third year of WP2 Task 1, specifically:

1. **T1.5 Self-directed search** where the aim is to build a Pattern Discovery (PD) module equipped with self-directed search. Also, this subtask included work on defining a set of segmental quality measures that was already addressed in Year 1 (M12) of the project. The expected completion date of milestone M2.1.5 "Self-directed search" was M30.

Task 2 - Pattern discovery with computational mechanics approaches

- To complete the deliverables and milestones that were deferred from the project's second year of operation pertaining to task T2.2, specifically the sections that deal with the study of **T2.2. Learning different hierarchical units from acoustic patterns**. This includes milestone M2.2.2A that was deferred at the beginning of Y3 from M15 to an expected completion date of M28. Also, another part of subtask T2.2 that deals with **Learning syntactic structure** was to be studied and includes milestone M2.2.2B. The latter was scheduled for completion in M27 but was moved to M32 at the beginning of year 3.
- To complete the deliverable and milestone that were set out for the third year of WP2 task 2 operation dealing with **T2.3 CMM Learning in ACORNS**. This includes milestone M2.2.3 whose completion date was extended from M33 to M36 at the beginning of Y3.

3.2.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved

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

During WP2's task 1 work period for the third year, efforts were concentrated on continuing the research undertaken during the first and second years related to pattern discovery, and extending them to support the overall goals of the ACORNS project. Specifically, research performed has covered the following areas where the contractor involved has been TKK unless otherwise specified.

- A novel permutation transformation method has been studied along with other ordinal methods for audio signal analysis and recognition. Ordinal methods operate in the time domain and are computationally lightweight as well as flexible. These methods are attractive from the ACORNS viewpoint since they collect information from the internal structure of signals.
- The development of a permutation transition frequency matrix for representing audio events was undertaken. The permutation transition frequency matrix can be considered as novel modelling tool for audio events, or for any time series events in general.

ACORNS

- The permutation transformation method was applied to the problem of stop-consonant recognition in the context of a vowel. Models in classification tests were derived from the consonant release part only and information from the upcoming vowel were not used in model generation.
- The results of the studies related to the permutation transformation were reported in the form of a presentation at the project's quarterly meeting in June 2009 in Helsinki as well as on the ACORNS project Wiki.
- Year 3 experiments common to the entire ACORNS consortium have been carried out by WP2. These experiments have included the integration and benchmarking of WP1 features.
- A study concerning the automatic discovery of sub-word units was performed and reported in WP2's Year 3 deliverable.
- The refinement and theoretical formulation of the Concept Matrix (CM) method reported in Year 2 was carried out. This included numerous experiments.
- A study concerning the applicability of different statistical and information theoretic measures for modeling of discrete time series was performed.
- WP2 worked on the development, testing and documentation of the Self-Learning Vector Quantization (SLVQ) software module. This software was made available to the other partners on the project's Wiki site during September and October of 2009.
- Collaboration with WP3 (Sheffield), WP4 (Leuven), and WP5 (Nijmegen) was carried out by working on joint publications.
- Concerted effort has been expended on dissemination by writing a number of publications which are listed in Deliverable D6.3

The activities described above have all been largely motivated by fulfilling the responsibilities related to the third year WP2 deliverable as well as the targeted milestone specified in the ACORNS Technical Annex (TA). We now describe the relationship of the above activities to the specific WP2 Task 1 goals outlined in the TA.

Milestone M2.1.5

For the completion of milestone M2.1.5 entitled "*Self-directed search*" work dealing with the construction of an automatic pattern discoverer equipped with self-directed search was required. Studies concerning this theme were conducted and completed by M30. This work culminated most directly with the publication of two conference papers, "A noise robust method for pattern discovery in quantized time series: the concept matrix approach", and "Self-learning Vector Quantization for Pattern Discovery from Speech", both presented at Interspeech '09, Brighton, UK. These papers were also published on the ACORNS project Wiki.

Task 2 - Pattern discovery with computational mechanics approaches

During the third year, WP2 Task 2 concentrated on continuing the research undertaken during the first two years related to CMM and extending it to support the overall goals of the ACORNS project. Specifically, research conducted during Year 3 has covered the following areas: (In the following list, the contractor involved for Task 2 activities has been TKK unless otherwise specified. The work on CMM theory and the modified CSSR algorithm has been performed by Gustav Henter located at KTH.)

Deliverable D0.3

ACORNS

- Work centred around the *causal state splitting reconstruction algorithm* (CSSR), a method showing promise for unsupervised pattern discovery, was continued.
- The limitations of CSSR discovered during the first two years of the project led to an extension of CSSR theory that is able to handle the effects of sequences containing noise. Work in this algorithm, called *CSSR with resolution*, was continued. During this study it was found that smaller representations that do not distinguish all causal states could achieve better performance for realistic sample sizes.
- Robust homogenization combined with regular determinization found in CSSR formed the basis for a new algorithm entitled *robust causal state discovery* (RCS). Besides developing the algorithm, theory was developed that is able to prove that RCS can recover the finite suffix partitioning of the underlying process, even when the sequence under observation is disturbed by noise up to a certain limit.
- Documenting the characteristics of the RCS algorithm has been carried out.
- A C++ implementation of RCS has been implemented and will be released as open source software under GPL.
- The results of RCS regarding learnability and robust causal state discovery are being formulated as a paper that will be disseminated within the academic community in the near future.

The Task 2 activities described above have been motivated by fulfilling the responsibilities related to the WP2 deliverables as well as the targeted milestones specified in the ACORNS TA. In the following section, we describe the relationship of the above activities to the specific WP2 Task 2 goals outlined in the TA.

CSSR, *CSSR with resolution*, as well as *RCS* software developed by Gustav Henter located at KTH was made available to the ACORNS project.

Milestone M2.2.2A

WP2 Task 2 was scheduled to work on **Learning different hierarchical units from acoustic patterns**. This milestone's completion time was extended to M28 at the beginning of Year 3. Up to now, results with CSSR-based algorithms have indicated that a limited cardinality of the symbol alphabet is required suggesting that the phone and phoneme level found in speech would be most applicable. However, no definite conclusion can yet be formed as to the overall applicability of CSSR to discover different hierarchical units from acoustic patterns due to the adverse effects of noise. For this reason work on robust homogenization was carried out so that sequences containing noise could be processed.

Milestone M2.2.2B

WP2 Task 2 was also scheduled to work on **Learning syntactic structure**. However, for reasons identical to those found for M2.2.2A above, work on learning syntactic structure was not undertaken.

Milestone M2.2.3

In WP2, task 2.3 dealt with **CMM learning in ACORNS**. This milestone was met in M36 as planned by reporting the current state of theoretic and algorithmic development pertaining to CSSR and RCS in WP2's year 3 deliverable. Also, an implementation of the algorithm was made available to the ACORNS project members, and will be published as open-source software. A journal article is being prepared as well.

3.2.3 Deviations from the project work programme, and corrective actions taken/suggested

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

The planned milestone for WP2 Task 1's third year of operation was met.

Task 2 - Pattern discovery with computational mechanics approaches

During the 2nd year's project review meeting in January 2009 held in Leuven, it was agreed with the project reviewers that both M2.2.2A and M2.2.2B would become *optional milestones*, to be fulfilled only if progress modifying the original CSSR algorithm to handle noise effectively was fully solved. Currently, efforts to manage noise effectively in CSSR-based algorithms is still underway, e.g., the *CSSR with resolution* algorithm has now evolved into a *robust causal state discovery* (RCS) form, and for these reasons the optional milestones were not reached. At the project's 2nd annual review meeting it was agreed that it would be more prudent to spend effort on developing a pattern discovery algorithm robust to noise, rather than spend time applying the CSSR algorithm - with its now known deficiencies to noisy sequences discovered within the ACORNS project - to different problems of speech decoding. Hopefully learning hierarchical units from acoustic patterns as well as learning syntactic structure can be addressed in the future.

3.2.4 List of Deliverables

Table 3.2.1: Deliverables List for WP2

Del. no.	Deliverable name	WP no.	Date due	Actual/Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months *)	Lead contractor
D2.1	Task 1: PD/DME Modules.	2	M12	M12			TKK
	Task 2: Applicability of CMM Learning for ACORNS		M9	M24			
D2.2	Task 1: Enhanced PD, Compact Coding, Linked DMEs	2	M24	M24			TKK
	Task 2: Learning linguistic units & syntactic structure with CMMs		M24	M36			

ACORNS

D2.3	Task 1: PD-module with self-directed search, derived segmental quality measures	2	M36	M36			TKK
	Task 2: Full integration of CMM to ACORNS		M36	M36			

*) if available

3.2.5 List of Milestones

Table 3.2.2: Milestones List for WP2

Milestone no.	Milestone name	WP no.	Date due	Actual/Forecast delivery date	Lead contractor
Task 1					
M2.1.1B	Auditory pre-processing with DMEs	2	M12 => M15	M15	TKK
M2.1.2	Enhanced PD for higher-level processing	2	M18	M18	TKK
M2.1.3	Temporal structures	2	M24	M24	TKK
M2.1.4	Auditory memory traces	2	M24	M24	TKK
M2.1.5	Self-directed search	2	M30	M30	TKK
Task 2					
M2.2.1	Applicability of CMM learning for ACORNS	2	M9	M24	TKK
M2.2.2A	OPTIONAL: <i>Learning different hierarchical units from acoustic patterns</i>	2	M15	Not applicable since milestone changed to <i>optional</i> in January 2009 at year 2 review meeting.	TKK
M2.2.2B	OPTIONAL: <i>Learning syntactic structure</i>	2	M27	Not applicable since milestone changed to <i>optional</i> in January 2009 at year 2 review meeting.	TKK
M2.2.3	CMM learning in ACORNS.	2	M33	M36	TKK

3.3 WP3 Memory Organisation and Access

3.3.1 Workpackage objectives and starting point of work at beginning of reporting period

The objectives of WP3 were:

- Creation of a synthetic memory structure that exhibits recognisable psychological behaviour as an emergent bi-product.
- Design and implementation of mechanisms and computational models of working memory architecture and access.
- Inclusion of attention within the overall memory architecture.
- Investigation and implementation of episodic and semantic memory within the memory-prediction Framework.
- Linking memory-prediction framework with dual-purpose sensory-motor representations

The starting point for period 3 of WP3 was the ACORNS memory architecture that includes evidence from psychological and emergent behaviour, and offers an interactive approach between attention, working memory and long-term memory. Various initial models were developed in period 2 that were directed by this architecture. The results from these ACORNS memory architecture models can be found in the WP report WP3.2.

3.3.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved

To reach the targets in Workpackage 3 described above, the following activities have taken place:

1. The memory architecture that was developed in the first two periods of the project was used as an overall framework for the development of speech recognition models.
2. Those models that been developed, inspired by the ACORNS memory architecture, include mechanisms that are based on the working memory component. For instance, the restricted Restricted Boltzmann Machine for spoken word recognition includes a working memory mechanism that combines learned weights and activations, at various levels of the hierarchical model, from the current input to produce representations. In the case of the Acoustic DP-Ngrams approach related to learning keywords working memory represents the acoustic and pseudo-visual modality for the current input during training and for recognition receives from episodic memory the appropriate pseudo-visual representation for the current speech input.
3. We have two main attention mechanisms inspired and constrained by the ACORNS memory architecture. The first provides special attentional focus and the second classifies speech from non-speech to gate what is introduced into working memory.

ACORNS

4. We have developed two main models that incorporate episodic long-term memory that perform keyword learning and word recognition. There as also been the development of a group of semantic long-term memory approaches that perform word recognition, associate semantic feature and speech signals, and early vocabulary acquisition. As such it as been possible to investigate the impact of these two memory structure from the ACORNS memory architecture on the types of models developed .
5. Consideration has been made with regards the association of sensor-motor representations.

3.3.3 Deviations form the project work programme, and corrective actions taken/suggested

There was no major deviation from the planned work programme.

3.3.4 List of Deliverable

Table 3.3.1: Deliverables List for WP3

Del. no.	Deliverable name	Work package 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
D3.2	Report focusing on the results of the initial ASR experiments comparing episodic and semantic memory	3	M24	26/11/08	28	28	USFD
D3.3	Report consolidating all of the results pertaining to memory organisation and access derived in WP3	3	M36	26/11/09	20	20	USFD

*) if available

3.3.5 List of Milestones

Table 3.3.2: Milestones List for WP3

Milestone no.	Milestone name	Workpackage no.	Date due	Actual/Forecast delivery date	Lead contractor
3.5	Final release of the software simulation of working memory (task 3.2) and attention mechanisms (task 3.3)	3	M30	30/07/2008	USFD

ACORNS

3.6	Final results of ASR experiments comparing episodic and semantic memory	3	M36	30/11/2008	USFD
-----	-------------------------------------------------------------------------	---	-----	------------	------

3.4 WP 4 Information discovery and integration

3.4.1 Workpackage objectives and starting point of work at beginning of reporting period

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.

The work in this period starts from the following results from the previous reporting periods:

- the Non-negative Matrix Factorisation (NMF)-based information discovery and integration technique that was developed in year 1 and 2 and that already achieved objectives 1 and 3
- the foundations of an exemplar-based speech recognition architecture which was described in the reports of the 2nd year.

3.4.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved

The work on NMF elaborates on objectives 1 and 3 by orienting it towards robustness and scalability.

Firstly, to address robustness we addressed non-ideal visual information. Since this overlaps with the objectives of WP5, we report on this work in D5.3.

Secondly, we realize that the vector quantization (VQ) that is currently part of the processing in the NMF framework may lead to loss of accuracy. We therefore searched for representations of speech that avoid this step or limit its accuracy loss:

- We applied the NMF algorithm to discover recurring patterns in the time-frequency plane. This is tantamount to a data-driven (bottom-up) way of designing a feature representation of speech. Since it also addresses the objectives of WP1, the description of this work, including the description of how this learned speech representation should be integrated in the NMF word discovery, grounding and recognition framework can be accomplished, was described in chapter 5 of D1.3. Originally, the plan was to use the activations of the time-frequency patterns as the input for word discovery with NMF. It turned out that including a VQ-step was still beneficial
- One way to alleviate information loss due to VQ could be to increase the codebook size, such that the quantization error is minimized. However, since the NMF is based on co-occurrence (bigram) statistics, the data requirements are quadratic in the codebook size, which is detrimental for the learning rate. Instead, we first applied VQ with different codebooks to the same data stream. The codebooks have different Voronoi regions, such that each data point can be located with greater accuracy by combining the information from the different codebooks. We observed that the error rate decreased with increasing number of codebooks. However, the accuracy leveled off around 5 to 10 codebooks. In a second experiment, we found that applying soft-VQ as proposed in (Van hamme, in Interspeech 2008) is a better compromise of required memory versus accuracy.
- Finally, we looked at consonant classification and estimation of articulatory features on the VCV-corpus (Cooke and Scharenborg, in Interspeech 2008) to see if greater time resolutions could improve the accuracy on this task in the NMF-framework. We found that window lengths of around 20ms are a good compromise on all tasks, including detection of plosives, voicing, place of articulation and overall consonant recognition. Higher time resolution does not seem to help in this test.

Thirdly, we examined methods for feature selection in the NMF framework with the goal of removing or downweighting noisy or uninformative co-occurrence event counts. This could be beneficial for:

1. robustness and accuracy. As shown in D5.3, adding irrelevant (visual) features decreases the NMF performance. By automatically downscaling the weight of features that do not support the learning process, irrelevant features could be removed. Also acoustic features that relate to irrelevant acoustic events such as silence could be removed.
2. self-organizing hierarchies. Suppose a first layer blindly (i.e. without supervision) discovers recurring acoustic patterns in the time-frequency plane. Now consider a second NMF layer which discovers words from visual information, the activations of the first layer and the inputs to the first layer. If feature selection (weighting) on the second layer now reveals that the inputs to the first layer are still useful for vocabulary acquisition, this implies that the learning of the first layer is not complete and its number of internal representations should be increased.

The year 3 work on the exemplar-based (also called episodic) speech recognition architecture in WP4 focused on objectives 2 and 4. A full report describing the system, its performance w.r.t. state-of-the art HMM systems, and its properties and relations with other methods explored in ACORNS can be found in deliverable D4.3.

The association of acoustic features and speech events (objective 3) in the exemplar-based approach is straight-forward since this merely requires that each piece of audio (frame) is labeled. Automatic discovery of the labels (objective 1) is abstracted in the work reported here. As is explained in deliverable D4.3, making a complete self-learning system requires the incorporation of either the NMF techniques explored in this work package, or the spectral clustering techniques¹ referred explored in year 2 (Master thesis Joost De Tollenaere, Milestone 4.2.2) or the DP-N-gram technique reported on in deliverables D2.3 and D3.3.

The inference engine (objective 4) forms the core of the exemplar-based system. The simple k-NN (nearest neighbor) based likelihood estimator used in the initial system was improved in several ways:

- The single frame (enriched with meta-information such as phonetic class label, gender of the speaker, or word) that served as basic unit in the k-NN computations was replaced by short traces, i.e. a short sequence of frames of a fixed and pre-defined length carrying the same meta-information as the central frame. The use of traces as basic units in the k-NN computation improves the inference since it automatically takes the acoustic context into account. The optimal length of the traces varied between 90msec (approximately one phone) and 250msec (three phones) depending on how much data is available on average for the distinct acoustic contexts.
- The distance measure for the k-NN estimator was also improved by using local sensitivity matrices derived from per-phone covariance matrices (see D4.3). Alternatively, the work of WP1 can also provide such a local sensitivity measure, based on an auditory model of human speech perception.
- The simple k-NN voting based likelihood estimator was replaced with a likelihood estimator based on local class-specific kernels. This new estimator adjusts the likelihoods based on the location of the local gravity center of the classes w.r.t. the input: the weight given to examples from a class for which the near examples for the current input are skewed to one side (instead of nicely surrounding the input) is lowered.
- As is postulated in the paper "The Case for Case-Based Automatic Speech Recognition" (see also deliverable D3.3), a key to more robust and well-performing inference may lay in making more "intelligent" use of the expert knowledge (typically of procedural nature) underlying "cognitive architectures". The k-NN system itself is underpinned by the general principle (knowledge) that similar observations tend to result in similar outcomes. Another generally applicable principle is that of continuity: most phenomena are continuous, the frequency of discontinuities (e.g. speaker changes, the slamming of a door) is substantially lower. A direct implementation (2D Viterbi decoding) of most continuity constraints is extremely costly even after having simplified the

¹ spectral clustering is a clustering technique based on matrix-factorization.

ACORNS

interdependencies to a minimum. Loopy belief propagation offers a viable alternative since it allows integration of such interdependencies with low overhead. Loopy belief propagation can hence be used to impose constraints on for example gender changes (never occurs within a phone or word) or to impute missing data in noisy recordings.

Keeping the computational burden within bounds when working with large data sets is of the utmost importance if exemplar-based systems are ever to be used for large vocabulary speech recognition. Vector processing and massive multi-threading were used to keep the turn-around time of experiments with the development system within practical bounds. The roadmap algorithm (a generalization of tree-based methods) is used as a fast (approximate) method for finding near matches of data with acoustic traces. An investigation on both artificial and real speech data showed that the roadmap algorithm can cope efficiently (memory & computation time) with large amounts of high dimensional data (the multiple frames that make up a trace). The roadmap algorithm forms the basis of a robust content addressable memory that can cope with large amounts of potentially incomplete (fuzzy) high dimensional data and serves as the implementation of Milestone M4.1.3. The intelligent interconnection of the memory cells (data points) with their 'interesting' neighbors allows for a quick traversal of the data space, and hence allows for quick access to the memory given new input data. The fact that a cell is either directly or indirectly connected with all surrounding cells also allows for checking a hypothesis (verification) by checking the consistency with the surrounding cells. This verification strategy forms the basis of the above mentioned 'local class-specific kernel' approach: information on the larger units higher up in the memory architecture (e.g. phone or word labels) is gathered from the surrounding cells and is used to adjust the estimated class likelihoods. In a second stage, belief propagation is used to add (impute) hypotheses on the higher level units, after which the k-NN lists can again be updated using the same roadmap algorithm, hence narrowing down the number of potential matches. An investigation on both artificial and real speech data showed that the roadmap algorithm can cope efficiently (memory & computation time) with large amounts of high dimensional data (the multiple frames that make up a trace). The experiments on the TIMIT database using the 'local class-specific kernel' approach showed the potential of the verification strategy.

The development system, only including a subset of the extensions presented above, was evaluated on both the ACORNS year 2 Dutch database and on the TIMIT database (a database used in WP1 and WP2 as well). The system showed results that are competitive with state-of-the-art HMM systems trained specifically for this task. On the ACORNS database, characterized by having a large set of examples for each acoustic condition, the example-based system even outperformed the HMM system. The evaluation on the TIMIT database also showed that the proposed refinements to the inference engine help when having only few examples per distinct acoustic condition.

3.4.3 Deviations from the project work programme, and corrective actions taken/suggested

The main focus in the work on the exemplar-based system was on the basic components that make-up the system (inference engine, roadmap, belief propagation) and making sure that the components can scale-up to large vocabulary recognition. The work on the inference engine took a lot more time than expected such that we did not succeed in integrating this scalable engine with the learning techniques of WP2 and WP4.

3.4.4 List of Deliverables

Table 3.4.1: Deliverables List for WP4

Del. no.	Deliverable name	WP 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	M24	22	15.5	KUL
D4.2	Report on LSA representation and SVD dimension reduction	WP 4	M24	M12	22	7	KUL
D4.3	Report on exemplar-based and activation-based matching	WP 4	M36	M36	22	18.6	KUL

*) if available

3.4.5 List of Milestones

Table 3.4.2: Milestones List for WP4

The milestone descriptions and target dates are the modified ones presented in the Periodic Activity Report of year 1.

Milestone No.	Milestone name	Workpackage no.	Date due	Actual/Forecast delivery date	Lead contractor
M4.1.1	Implementation of activation mechanisms at the tag level.	WP 4	M12	M12	KUL
M4.1.2	Top-down learning of patterns and computation of activations	WP 4	M24	M24	KUL
M4.1.3	Activation/verification mechanisms in hierarchical CAMs	WP 4	M36	M36	KUL
M4.2.1	LSA representation and SVD dimension reduction	WP 4	M12	M12	KUL
M4.2.2	ASMs defined from WP1 and WP2 features and automatic segmentation	WP 4	M36	M24	KUL
M4.3.1	Time synchronous exemplar-based and activation based matching	WP 4	M24	M24	KUL

M4.3.2	Time-asynchronous matching and non-Euclidean distance	WP 4	M36	M36	KUL
--------	-------------------------------------------------------	------	-----	-----	-----

3.5 WP5 Interaction and communication

3.5.1 Workpackage objectives and starting point of work at beginning of reporting period

The work package objectives, as defined in the TA, are:

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. We will provide 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. At 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 it 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.

At the start of the reporting period, we could take the recommendations of the Scientific Advisory Committee and the Second Review as starting point. Below, we indicate how we addressed the recommendations from the reviewers (to the extent relevant for WP5), in a point-by-point manner.

Recommendation 1

We recommend that no new approach be explored during the last year, but rather to concentrate on the most promising ones and to elaborate a detailed qualitative and quantitative comparison among these approaches

We have followed this recommendation by doing experiments that compare and contrast the different approaches for comparable tasks. Our intention was to investigate the strengths and weaknesses of the approaches, rather than ranking them along a conventional scale defined by 'keyword error rate'. On the basis of all experiments, our conclusion is that the performance of any approach to a large extent depends on the subtle assumptions that are underlying the conceptual choices made in these computational approaches. The experiments and the conclusions have been described in D5.3.

Recommendation 2

Deliverable D0.3

ACORNS

We recommend that a unique well-defined experimental scenario, together with a unique database, be used as a basis (which does not prevent using other scenarios and databases to further evaluate individual methods)

In the Interspeech paper (ten Bosch et al., 2009), all computational methods are compared on the same training and test data. This paper has been included in D5.3. In other experiments, described in D1.3 and also in D5.3, methods have been subject to a systematic exploration of the various experimental factors including noise type, noise level, preprocessing, and channel normalisation. In all experiments, the starting point was a carefully designed common database with different but comparable training sets and independent test sets.

The ACORNS interaction platform will be made available via a public deliverable of all software. The entire ACORNS speech database (4 languages, about 35 speakers, about 50000 utterances in total, including orthographic annotation and tag metadata) will be made available by a separate deliverable.

Recommendation 3

We recommend that the new features developed in WP1 be reused to run all the experiments performed in years 1 and 2 with more classical features, and observe whether it makes a difference or not in terms of performances;

In WP1 and in WP5, experiments have been done to see what novel features (features that deviate from the conventional MFCC features) bring. The experiments show that the modified MFCCs (MMFCCs) and static-adaptive MMFCCs (SA-MMFCCs) show improvements for specific (classical ASR-type) experiments – however, not all improvements could be generalized to the ACORNS experiments. These results have been carried out in October and November 2009. The results were particularly interesting because they again showed the importance of the back-end learning/decoding mechanism for the optimization of features. The results strongly suggest that for the improvement of learning results issues concerning correlation and non-normal distribution of features are playing a major role. The experiments have been described in D1.3 and D5.3.

3.5.2 Progress towards objectives – tasks worked on and achievements made with reference to the planned objectives, identify contractors involved

Task 5.1 Creation of a platform for learning in the memory-prediction framework

In the ACORNS project, a software platform has been developed that is used for relating specialized MATLAB modules (developed by ACORNS partners) into an overall caregiver-learner interaction framework. The platform is part of deliverable D6.3, which includes a zip file with software and documentation. The platform includes the caregiver-learner interaction, the three major computational learning approaches NMF, DP-Ngrams, and Concept Matrices, and information about a number of other computational approaches such as the State Splitting and Recombination algorithm. The platform will be available as open source under the GNU GPL license.

The entire ACORNS software deliverable consists of the following packages:

- CA-LA-interaction (MATLAB)
- Concept Matrices and Segmentation (C)
- SCCR_RCSD (C)
- DP_Ngrams (MATLAB)
- Features (MATLAB)
- Interaction Platform (MATLAB)
- Self Learning VQ (MATLAB)

- NMF (MATLAB)

For more details, we refer to D6.3.

Task 5.2 Multimodal integration

In this task, the scientific question is how information from different sources in different streams can/must be combined for optimal learning/decoding performance. This task has been addressed by focussing how audio and visual information can be integrated, with particular focus on the type of encoding of the information available in the ‘visual’ stream. This encoding issue was already inspired by the experiments based on networks such as Self Organizing Maps (SOMs, Klein et al., 2008) and by word learning experiments involving multilingual input (these experiments took place in year 1). The basic question is to what extent the grounding information in the visual domain as used in year 1 and 2 was too ‘crisp’ and to what extent this information could be made less invariant and therefore more realistic. The follow-up question was what types of low level (‘sensory’) or high level (‘conceptual’) features could be used.

Deliverable D5.3 contains a section in which we explain how we can deviate from the invariant, symbolic tags (as used in the year 1 and part of year 2 experiments) and develop a more realistic and plausible encoding of the visual channel. Furthermore we explored various types of variation in this channel, which accommodates between-type and between-token variation. The main results are (a) the encoding of the visual channel is a complex, theoretically interesting issue with direct links on the representation of the physical word, and with relations to the concept of ‘word’ (b) the value of a certain encoding also critically depends on the particular mathematical properties of the decoding back-end.

More details are provided in D5.3

Task 5.3 Architecture for interaction

In the open source software deliverable, one of the packages directly deals with the caregiver-learner (caring agent = CA, learning agent = LA) interaction, abbreviated as CA-LA-interaction. The software in this package has explicitly been used in interaction experiments. One of the tasks in WP5 was studying how learning, more particular the learning results, depends on interaction between caregiver and learner. In most ACORNS experiments, we have assumed that the caregiver always present complete and consistent stimuli. Each stimulus consists of an audio part and a ‘visual’ (grounding) part. In these experiments, the learner takes each stimulus ‘as it appears’. That is, the learner assumes the stimulus to be consistent, and does not doubt the consistency between the modalities within the stimulus.

We have done experiments with the aim to investigate what would happen if the interaction between caregiver and learner was modified towards less strict forms of supervision during training. We investigated four different interaction strategies between caregiver and learner. This has been published in and presented at the Workshop for Child-Computer Interaction (WOCCI-09).

We also investigated another interesting phenomenon that may take place between caregiver and learner. If the learner is able to take a more active role (in the sense that the learner is able to *overrule* the information presented in the stimulus or *fill in* missing information in the stimulus), then the learner is able to recover from a certain level of inconsistency in the presented stimuli. In other words, a certain level of contrariness at the learner’s side helps to overcome inconsistencies in stimuli from the caregiver.

This is explained in more detail in D5.3. In this deliverable, we explain how the learner is endowed with a decision mechanism based on internal confidence measures.

Task 5.4 Experiments with language learning

One of the central aims of doing experiments was to relate model results to findings described in the literature on language acquisition. In this literature, many findings are reported – however, it is not always

ACORNS

straightforward to compare these findings with computational experiments on specific data sets. One of the findings, reported in Newman (2008), provided a good example of an empirical result that could be compared and contrasted with results obtained by computational simulation. Newman's statement is that young infants are better in recognizing novel speakers if they have been exposed to more different speakers earlier.

Such an experiment is an excellent opportunity to contrast the three computational approaches NMF, CM, and DP-Ngrams, which was one of the recommendations of the reviewers. The experiment is published in Interspeech 2009 (ten Bosch et al, 2009 – see www.acorns-project.org). The fact that the computational results are in line with Newman's observations inspired us to invite her as speaker on the second ACORNS workshop, as a satellite of Interspeech 2009 in Brighton.

In brief, the results show that methods such as CM and NMF show a better generalization to novel test speakers in the case of multi-speaker training condition than in the single-speaker training condition. For the computational approach DP-Ngrams, the difference between single-speaker and multi-speaker training condition is much smaller. We interpreted this as an interesting result, since it strongly suggests that DP-Ngrams, as the most episodic approach of the three, deals with generalisation to novel acoustic situations in a different way than the two other approaches.

Finally, WP5, together with WP1, investigated the question what would happen in the case of realistic noise using novel features in the case of language learning. This study mainly relates to WP1 (features) in combination with WP5 (experiments). Since the computational approaches (NMF, CM, DP-Ngrams) differ with respect to how episodic information is dealt with during the learning, it is expected that they differ with respect to their robustness against background noise. It was expected that the novel features that were designed in WP1 show better performance in adverse conditions. WP1 designed two novel types of features ('modified'-MFCC and 'static adaptive'-MFCC); both these features were subject to various tests. Experiments showed that MMFCC outperforms MFCC in both the 'classical' ASR-based experiments and in ACORNS experiments. The static-adaptive MMFCCs showed improvements in HMM-ASR-based tests, but not in the ACORNS learning experiments, showing that improvements in features depend on the decoding back-end.

More details are provided in D1.3 and D5.3.

3.5.4 Deviations from the project work programme, and corrective actions taken/suggested

The single main change in WP 5 was based on recommendations by the Scientific Advisory Committee (SAC, end of 2008) and the second review (Jan 2009), not to go for 'ever more words' but instead to focus on gaining insights from experiments and to put emphasis on understanding the learning procedures. This is explained in more detail in D5.3.

This recommendation was actually welcomed and supported by all members of the consortium, since it opened the way to many interesting learning experiments, as evidenced in the ACORNS publications and deliverables.

3.5.4 List of Deliverables

Table 3.5.1: Deliverables List for WP5

Del. no.	Deliverable name	WP 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	WP 5	M12 = 30 Nov 2007	M12 = 30 Nov 2007			RUN
D5.2	System capable of learning a 50 word vocabulary	WP 5	M24 = 30 Nov 2008	M24 = 30 Nov 2008			RUN
D5.3	System capable of rapidly learning a large vocabulary **)	WP 5	M36 = 30 Nov 2009	M36 = Nov 30, 2009			RUN

*) if available **) This is the original title according to the TA. The contents of the deliverable is adapted based on SAC and second review recommendations.

3.5.5 List of Milestones

Table 3.5.2: Milestones List

Milestone no.	Milestone name	WP no.	Date due	Actual/Forecast delivery date	Lead contractor
M5.4	Specification of second year experiments	5	M15	M18	RUN
M5.6	Specification of third year experiments	5	M27	M25-M28	RUN
M5.7	Complete implementation of final learning system	5	M34	M35	RUN
M5.8	Release of open source software for memory-prediction based learning	5	M34	M35	RUN

3.6 WP 6 dissemination and Use

There are five tasks in this workpackage.

The first task is related to the maintenance of a public website. The project website was regularly updated, see www.acorns-project.org. The public website provides access to all publications and public deliverables. In addition, it provides general information, information on the workshops, contact information for the consortium and the consortium members, meeting dates, etc.

The second task relates to organising a workshop dedicated to topics of ACORNS. The final workshop took place on 11th of September 2009, as a satellite event to the large-scale Interspeech Conference in Brighton, UK. We have organised a one day workshop that had a number of goals. First, it presented the results of the ACORNS project. Second we brought together the top researchers from the different disciplines that are important for the topic of the ACORNS project: Deb Roy (modelling of language acquisition based on daily real life audio and video recordings), Friedeman Pulvermüller (cognitive and behavioural neuroscience), and Rochelle Newman (language acquisition). And third, discussions were held to prepare new research proposals.

The third task relates to open source software. Information about available software and speech corpora is provided in deliverable 6.3.

The fourth task deals with the publications in ACORNS. In the third year of ACORNS has resulted in many publications. There were 3 journal papers and 16 conference papers.

The fifth task deals with public awareness. In this context we had a presentation in the first FET conference in Prague, 21-23 April.

Journal Papers

1. Louis ten Bosch, Hugo Van hamme , Lou Boves, Roger K. Moore "A computational model of language acquisition: the emergence of words", *Fundamenta Informaticae*, Vol. 90, (2009), pp. 229-249.
2. Maarten Van Segbroeck and Hugo Van hamme "Unsupervised learning of time-frequency patches as a noise-robust representation of speech", *Speech Communication*, Vol.51, (2009), pp.1124-1138
3. Veronique Stouten and Hugo Van hamme "Automatic voice onset time estimation from reassignment spectra", *Speech Communication*, Vol. 51, (2009), pp. 1194-1205.

Conference Papers

5. Guillaume Aimetti "Modelling Early Language Acquisition Skills: Towards a General Statistical Learning Mechanism", *Proc. EACL-2009*.
6. Okko Räsänen & Joris Driesen "A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition", *Proc. 17th Nordic Conference on Computational Linguistics*, 2009.
7. Louis ten Bosch, Joris Driesen, Hugo Van hamme, Lou Boves "On a computational model for language acquisition: modeling cross-speaker generalisation", *Proc. Text, Speech and Dialogue, 12th Intern. Conference, TSD 2009*.

ACORNS

8. Guillaume Aimetti, Roger K. Moore, Louis ten Bosch, Okko Räsänen, Unto K. Laine "Discovering Keywords from Cross-Modal Input: Ecological vs. Engineering Methods for Enhancing Acoustic Repetitions", *Proc. Interspeech 2009*.
9. Roger K. Moore, Louis ten Bosch "Modelling Vocabulary Growth from Birth to Young Adulthood", *Proc. Interspeech 2009*.
10. Okko J. Räsänen, Unto K. Laine, Toomas Altsaar "A noise robust method for pattern discovery in quantized time series: the concept matrix approach", *Proc. Interspeech 2009*.
11. Okko J. Räsänen, Unto K. Laine, Toomas Altsaar "An Improved Speech Segmentation Quality Measure: the R-value", *Proc. Interspeech 2009*.
12. Okko J. Räsänen, Unto K. Laine, Toomas Altsaar "Self-learning Vector Quantization for Pattern Discovery from Speech", *Proc. Interspeech 2009*.
13. Viktoria Maier, Roger K. Moore "The Case for Case-Based Automatic Speech Recognition", *Proc. Interspeech 2009*.
14. Saikat Chatterjee, Christos Koniaris and W. Bastiaan Kleijn "Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance", *Proc. Interspeech 2009*.
15. L. ten Bosch, O. Räsänen, J. Driesen, G. Aimetti, T. Altsaar, L. Boves, A. Corns "Do Multiple Caregivers Speed up Language Acquisition?" *Proc. Interspeech 2009*.
16. Joris Driesen, Louis ten Bosch, Hugo Van hamme "Adaptive Non-negative Matrix Factorization in a Computational Model of Language Acquisition", *Proc. Interspeech 2009*.
17. Mark Elshaw, Roger K. Moore "A recurrent working memory architecture for emergent speech representation", *The Bernstein Conference on Computational Neuroscience (BCCN)*, 2009.
18. Mark Elshaw, Roger K. Moore and Michael Klein "Hierarchical recurrent self-organising memory (H-RSOM) architecture for an emergent speech representation towards robot grounding", *Proc. Conference on Natural Computing and Intelligent Robotics*.
19. Louis ten Bosch, Lou Boves and Okko Rasanen "Learning meaningful units from multimodal input – the effect of interaction strategies", *Proc. Wocci2009*.
20. Guillaume Aimetti, Louis ten Bosch, Roger K. Moore "The emergence of words: Modelling early language acquisition with a dynamic systems perspective", *Proc. EpiRob-09.*]

Theses

All publications can be accessed through the public website maintained by the project.

Presentations about the ACORNS project

Task five is devoted to spreading awareness beyond the scientific community.

Table 3.6.1: Deliverables List

Del. no.	Deliverable name	WP no.	Date due	Actual/ Forecast delivery date	Estimated indicative person-months *)	Used indicative person-months	Lead contractor
D6.2.1	First Project Workshop	6	M18	M12		1/2	RUN
D6.2.2	Second Project Workshop	6	M36	M33		1/2	RUN

ACORNS

D6.4	Published papers	6	M24	M24		1/2	RUN
D6.5	Public Awareness	6	M36	M36			RUN

- 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 second year were somewhat harder than for the first year, since attention had to be paid to better harmonize and integrate the work of the workpackages. All partners were willing to work on better integration, but it took some effort to accomplish this. We decided to have two and a half days project meetings (instead of 1 and a half days), so that there would be more room for in-depth discussion on specific topics.

The meetings were organised as planned, the minutes of the meetings and audio conferences were always sent in time.

In order to guarantee and optimise the quality of the deliverables we have assigned two or three senior members of the consortium –who were not involved in producing a specific deliverable- to review the drafts of the texts.

- **Contractors**

All partners have enthusiastic and dedicated teams working on the project. Only KTH was struggling to get the right persons in time. The senior staff members still spend relatively much time to the project. 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.

– **Project timetable and status**

WP	Task	Month	25	26	27	28	29	30	31	32	33	34	35	36
		Task												
WP0	Project Management	T0.1			Q			Q			Q			D0.2.3 D0.3.3
		T0.2												D0.4.3
WP1	Signal Representations	T1.1						M1.4						D1.3
		T1.2												
WP2	Signal Patterning	T2.1						M2.1.5						M2.1B D2.3
		T2.2			M2.2B									
WP3	Memory Organization and Access	T3.1						M3.3.2						D3.1 M3.2
		T3.2												
		T3.3												
		T3.4												
		T3.5												
WP4	Information Discovery & Integration	T4.1												D4.3 M4.2.2 M4.3.2
		T4.2												
		T4.3												
WP5	Integration and Communication	T5.1												M5.8 D5.3
		T5.2									M5.7			
		T5.3						M5.2						
		T5.4			M5.6									
WP6	Dissemination and Standardization	T6.1			D6.1									D6.2.2 D6.3 D6.4.3
		T6.2												
		T6.3												
		T6.4												
		T6.5												

Updated Gantt Chart for Year 3

Overall, the work has proceeded according to the original plans, as updated at the end of the first year.

– **Co-ordination activities**

Four project meetings took place during the third year. The first meeting was dedicated to the preparation of the second annual review. The next two meetings, which were fully dedicated to scientific information exchange and planning of research, took two and a half days, while the last meeting, dedicated to reviewing the final deliverables, was confined to two full days. In addition to the face-to-face meetings 5 telephone conferences were held, each taking about one hour.

Apart from the meetings at the level of the full consortium intensive communication between researchers took place using e-mail and telephone.

The following meetings took place:

- | | |
|---------------------------------------------------------|-------------------|
| 1. Preparation of the Review meeting in Leuven, Belgium | 26 January 2009 |
| 2. Project meeting in Berg en Dal, Netherlands | 11-13 March 2009 |
| 3. Project meeting in Espoo, Finland | 22-24 June 2009 |
| 4. Project meeting in Grindleford, UK | 2-4 November 2009 |

The second workshop took place on 11 September 2009, in Brighton, UK.

Audio conferences took place on:

- 29 January 2009
- 27 March
- 5 June 2009
- 17 August 2009
- 27 October 2009

Annex 1: Plan for dissemination and Use

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. Procedure for CM		1. Speech recognition	2010	patent application (FIN-20075696) filed	TKK
3. Software package for speech signal processing		1. speech recognition and speech coding	After 2010	n.a.	KTH
4. Structure detection by means of Non-Negative Matrix Factorisation		1. Speech recognition 2. Data mining	After 2010	n.a.	KU Leuven
5. Improved software implementation of the CSSR algorithm for Computational Mechanics Modelling		1. Data mining, 2. Structure discovery	2010	n.a.	KTH
6. Platform for conducting experiments with simulating language acquisition		Scientific research	2010	n.a.	RU and all partners
7. ACORNS Corpus					TKK and all partners
8. NWO project		Scientific research			RU

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

ACORNS

- 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 far 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.
 - Additional processing modules will be included to compute features that are especially salient from an auditory point of view
 - Features will be provided on millisecond, deci-second and centi-second time scales
 - 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 from physical processes, represented in terms of non-negative numbers (e.g. 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 original NMF algorithms have been adapted to enable incremental decomposition, equivalent to incremental learning.
 - 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.
4. Causal State Splitting and Reconstruction algorithm for Computational Mechanics Modelling
- New implementation of the algorithm, repairing small bugs in the publicly available implementation.
 - Extension of the algorithm to allow for approximate (instead of exact) causal states, to make the approach robust against natural variation in many kinds of data.
5. Platform for simulating language learning
- The platform consists of a number of MATLAB scripts that implement the learner, the care giver and the interaction between learner and care giver.
 - Stimuli can be formatted with different amounts of information about the situational context in which speech utterances can be interpreted.
 - The actions of the care giver include the selection of new stimuli to offer to the learner, the interpretation of the learner's response and the decision on how to proceed.
 - The actions of the learner depend on the learning algorithm(s) selected by the experimenter.

- The platform provides a choice of techniques for monitoring and interpreting the performance of the learner.

2 Dissemination of knowledge

Table 1.1 Overview table of past and future 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
26/11 – 28/11/2007	Workshop	Research	global	35	RU Nijmegen and USFD
11-9-2009	Workshop Interspeech-2009 Satellite event	Research	global	50	RU Nijmegen
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
24/10/2008	Proposal for ESF Research Network	Scientific	Europe	thousands	RU Nijmegen

No specific dissemination activities were planned for the reporting period. However, a special session in the Interspeech conference in Brisbane was dominated by papers from the ACORNS project.

The project website (<http://www.acorns-project.org>) has been extended and maintained during the reporting period. The website is being kept up-to-date by the project coordinator.

The website will be maintained and updated for at least three years after the conclusion of the project. This will enable the project to provide access to publications that will see the light only after the formal end of the contract.

In October 2008 a proposal for an ESF Research Network “Language models of Evolution, Acquisition, and Processing” has been submitted, as a follow-up action to anchor the results of the research in ACORNS.

List of publications

Year-1

Papers

Lou Boves, Louis ten Bosch, Roger Moore "ACORNS -- towards computational modeling of communication and recognition skills", Proc. ICCI-2007.

Veronique Stouten, Kris Demuyneck, Hugo Van hamme "Automatically Learning the Units of Speech by Non-negative Matrix Factorisation", Proc. Interspeech 2007.

Veronique Stouten, Kris Demuyneck, Hugo Van hamme "Discovering Phone Patterns in Spoken Utterances by Non-Negative Matrix Factorization", IEEE Signal Processing Letters 2008

ACORNS

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].

Year-2*Papers:*

Hugo Van hamme "Integration of Asynchronous Knowledge Sources in a Novel Speech Recognition Framework", ISCA ITRW, *Speech Analysis and Processing for Knowledge Discovery*

Louis ten Bosch, Hugo Van hamme, Lou Boves "Unsupervised detection of words – questioning the relevance of segmentation", ISCA ITRW, *Speech Analysis and Processing for Knowledge Discovery*

Louis ten Bosch, Lou Boves "Language acquisition: the emergence of words from multimodal input", in Sojka, P., Horák, A., Kopeček, I & Pala, K. (Eds.) *Text, Speech and Dialogue, 11th Intern. Conference, TSD 2008*, Brno, pp. 261-268

Klein, M., Frank, S., van Jaarsveld, H., ten Bosch, L.F.M., & Boves, L. "Unsupervised learning of conceptual representations - a computational neural model", *Proc. 14th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP)*, 4-6 September 2008, Cambridge, UK

Okko Räsänen, Altosaar, T. & Laine U.K. (2008) Comparison of prosodic features in Swedish and Finnish IDS/ADS speech. *Proc. of Nordic Prosody X*.

Okko Räsänen, Unto K. Laine, Toomas Altosaar "Computational language acquisition by statistical bottom-up processing", *Proc. Interspeech 2008*, pp. 1980-1983

Joris Driesen, Hugo Van hamme "Improving the Multigram Algorithm by using Lattices as Input", *Proc. Interspeech 2008*, pp. 2086-2089

Hugo Van hamme "HAC-models: a Novel Approach to Continuous Speech Recognition", *Proc. Interspeech 2008*, pp. 2554-2557

Joost van Doremalen, Lou Boves "Spoken Digit Recognition using a Hierarchical Temporal Memory", *Proc. Interspeech 2008*, pp. 2566-2569

Louis ten Bosch, Hugo Van hamme, Lou Boves "A computational model of language acquisition: focus on word discovery", *Proc. Interspeech 2008*, pp. 2570-2573

Book Chapter

Louis ten Bosch, Hugo Van hamme, Lou Boves "Discovery of words: Towards a computational model of language acquisition", in: France Mihelič and Janez Žibert (Eds.) *Speech Recognition: Technologies and Applications*, Vienna: I-Tech Education and Publishing KG, pp. 205 - 224

Theses

Joost van Doremalen "Hierarchical Temporal Memory Networks for Spoken Digit Recognition", Radboud University Nijmegen, Dept. of Language & Speech, December 2007.

Joost De Tollenaere "Zelflerende spraakherkenning: akoestische eenheden en woordmodellen" MSc thesis, K.U.Leuven, ESAT, 2008, (in Dutch)

Year 3

Journal Papers

- Louis ten Bosch, Hugo Van hamme , Lou Boves, Roger K. Moore "A computational model of language acquisition: the emergence of words", *Fundamenta Informaticae*, Vol. 90, (2009), pp. 229-249.
- Maarten Van Segbroeck and Hugo Van hamme "Unsupervised learning of time-frequency patches as a noise-robust representation of speech", *Speech Communication*, Vol.51, (2009), pp.1124-1138
- Veronique Stouten and Hugo Van hamme "Automatic voice onset time estimation from reassignment spectra", *Speech Communication*, Vol. 51, (2009), pp. 1194-1205.

Conference Papers

- Guillaume Aimetti "Modelling Early Language Acquisition Skills: Towards a General Statistical Learning Mechanism", *Proc. EACL-2009*.
- Okko Räsänen & Joris Driesen "A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition", *Proc. 17th Nordic Conference on Computational Linguistics*, 2009.
- Louis ten Bosch, Joris Driesen, Hugo Van hamme, Lou Boves "On a computational model for language acquisition: modeling cross-speaker generalisation", *Proc. Text, Speech and Dialogue, 12th Intern. Conference, TSD 2009*.
- Guillaume Aimetti, Roger K. Moore, Louis ten Bosch, Okko Räsänen, Unto K. Laine "Discovering Keywords from Cross-Modal Input: Ecological vs. Engineering Methods for Enhancing Acoustic Repetitions", *Proc. Interspeech 2009*.
- Roger K. Moore, Louis ten Bosch "Modelling Vocabulary Growth from Birth to Young Adulthood", *Proc. Interspeech 2009*.
- Okko J. Räsänen, Unto K. Laine, Toomas Altsaar "A noise robust method for pattern discovery in quantized time series: the concept matrix approach", *Proc. Interspeech 2009*.
- Okko J. Räsänen, Unto K. Laine, Toomas Altsaar "An Improved Speech Segmentation Quality Measure: the R-value", *Proc. Interspeech 2009*.
- Okko J. Räsänen, Unto K. Laine, Toomas Altsaar "Self-learning Vector Quantization for Pattern Discovery from Speech", *Proc. Interspeech 2009*.
- Viktoria Maier, Roger K. Moore "The Case for Case-Based Automatic Speech Recognition", *Proc. Interspeech 2009*.
- Saikat Chatterjee, Christos Koniaris and W. Bastiaan Kleijn "Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance", *Proc. Interspeech 2009*.
- L. ten Bosch, O. Räsänen, J. Driesen, G. Aimetti, T. Altsaar, L. Boves, A. Corns "Do Multiple Caregivers Speed up Language Acquisition?" *Proc. Interspeech 2009*.
- Joris Driesen, Louis ten Bosch, Hugo Van hamme "Adaptive Non-negative Matrix Factorization in a Computational Model of Language Acquisition", *Proc. Interspeech 2009*.
- Mark Elshaw, Roger K. Moore "A recurrent working memory architecture for emergent speech representation", *The Bernstein Conference on Computational Neuroscience (BCCN)*, 2009.
- Mark Elshaw, Roger K. Moore and Michael Klein "Hierarchical recurrent self-organising memory (H-RSOM) architecture for an emergent speech representation towards robot grounding", *Proc. Conference on Natural Computing and Intelligent Robotics*.
- Louis ten Bosch, Lou Boves and Okko Rasanen "Learning meaningful units from multimodal input – the effect of interaction strategies", *Proc. Wocci2009*.
- Guillaume Aimetti, Louis ten Bosch, Roger K. Moore "The emergence of words: Modelling early language acquisition with a dynamic systems perspective", *Proc. EpiRob-09*.]

All publications can be accessed through the public website maintained by the project.

Planned presentations and publications

Mark Elshaw, Roger K. Moore and Michael Klein “An attention-gating recurrent working memory architecture for emergent speech representation”, *Journal of Connection Science* (Accepted awaiting proof copy).

All partners, “A new memory based architecture for acquisition of communication and recognition skills”. *Journal of Neural Networks* (Awaiting resubmission)

All partners: an integrated paper on ACORNS results including final experiments on semi-supervised training. Target journal is *Computer Speech and Language*:

Laine U. & Räsänen O.: "Indirect estimation of formant frequencies through mean spectral variance with application to automatic gender recognition.", accepted for publication, 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), Firenze, Italy, 2009.

- Räsänen O., Laine U.K.: “A Method for noise robust context-aware pattern discovery from symbolic time series”, in preparation.

- Räsänen O., Laine U.K. & Altsaar T.: "Blind segmentation of speech using non-linear filtering methods, in preparation.

Papers covering the permutation transformation currently in preparation for future publishing include:

- Seppo Fagerlund, Unto Laine. *Stop Consonant Recognition using the Permutation Transformation*
- Seppo Fagerlund, Unto Laine. *Ordinal Methods in Time Series Analysis and Recognition*

5.3 Publishable results

No public domain software was planned to be released in the reporting period.

We consider making available a novel implementation of the CSSR algorithm for Computational Mechanics Modelling that was developed in the reporting period available in Open Source format.

Spin-off projects

VASI: Vocabulary Acquisition from Speech and Images.

October 2009 - September 2013

Research fund K.U.Leuven

Partners: Speech and image processing research groups of K.U.Leuven

Budget: 450kEURO

The goal of this project is to explore data-driven, weakly supervised and unsupervised methods for hierarchical object discovery in speech and image recognition. In contrast to the mainstream model-driven, highly supervised and fine-tuned schemes, we investigate what can be learned directly from data, in a bottom-up fashion, by discovering re-occurring patterns and structures at multiple levels of abstraction. The main challenge then consists of finding out which types of variations are relevant and which are not. Here, we hypothesize that patterns that are found consistently over multiple modalities are probably the semantically more meaningful and interesting ones. Hence integration of different modalities is key for such unsupervised or weakly supervised approaches to succeed. Linking utterances to visual perception and vice versa, we come closer to semantic-level image and speech understanding, learned much like in human learning.

Use-plan

The research outputs from WP3 have already formed the basis for two PhD theses, and is feeding out into parallel research in other areas. For example, at USFD the DP-Ngram approach is being investigated in the area of heart condition monitoring.