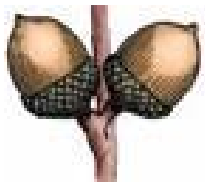

Project no. 034362

ACORNS

Acquisition of COmmunication and RecogNition Skills

Instrument: STREP
Thematic Priority: IST/FET

D3.3 Report consolidating all of the results pertaining to memory organisation and access derived in WP3

Due date of deliverable: 2009-11-30
Actual submission date: 2009-11-26

Start date of project: 2006-12-01 Duration: 36 Months

Organisation name of lead contractor for this deliverable: Speech and Hearing Research Group, University of Sheffield

Revision: 0.4

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

VERSION DETAILS
Version: 0.4 **Date: 26/11/09** **Status: Final**

CONTRIBUTOR(S) to DELIVERABLE	
Partner	***Name***
Speech and Hearing Research Group, University of Sheffield	**Guillaume Aimetti, Mark Elshaw, Robin Hofe , Vicky Maier, Roger K Moore**
Centre for Language and Speech Technology, Radboud University Nijmegen	**Michael Klein, Louis ten Bosch**
Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology	**Okko Räsänen**
Center for Processing Speech and Images, Katholieke Universiteit Leuven	**Joris Driesen, Hugo Van hamme**

DOCUMENT HISTORY			
Version	***Date***	***Responsible***	***Description***
01	**25/10/09**	**Mark Elshaw**	**First draft of report completed**
02	**26/10/09**	**Robin Hofe**	**Review content**
03	**10/11/09**	**Mark Elshaw and Robin Hofe**	**Restructure and add content based on reviewers comments**
04	**26/11/09**	**Mark Elshaw**	**Restructure based on comments of reviewers**

DELIVERABLE REVIEW			
Version	***Date***	***Reviewed by***	***Conclusion****
02	**27/10/09**	**Kris Demuynck**	**Restructure report and bring out comparison**
02	**27/10/09**	**Bastiaan Kleijn**	**Restructure**
03	**25/11/09**	**Kris Demuynck**	**Restructure report and include models in D4.3**
03	**25/11/09**	**Bastiaan Kleijn**	**Restructure report**

TABLE OF CONTENTS

1 Introduction

This WP3.3 report consolidates the findings related to the memory architecture produced as part of the ACORNS project. This consolidation brings together the findings by relating them to the WP3 objectives of:

- Creation of a synthetic memory structure that exhibits recognizable 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

Devised and developed over the ACORNS project, the ACORN memory architecture forms the basis for various memory models to achieve the objectives outlined above. Examining these models offers the opportunity to consider how combining various components of the memory architecture (sensor and motor working memory, semantic long-term memory, episodic long-term memory, motor long-term memory and attention) direct them. For speech acquisition this memory architecture combines attention, working memory and long-term memory. According to Burgess and Hitch (2005), many computational models of human memory concentrate on working memory or long-term memory at the expense of the other. However, the ACORNS memory architecture incorporates the interaction between working memory and long-term memory into a single architecture. The memory models developed within the memory architecture for speech acquisition split into three main areas: (i) attention based mechanisms, (ii) combined working memory and semantic long-term memory, and (iii) combined working memory and episodic long-term memory. In addition, there is a consideration in this report of a sensorimotor representation approach to incorporate speech production in the ACORNS memory architecture.

In this report, we will concentration on two episodic long-term memory models: the Acoustic DP-Ngram and Temporal Episodic Memory Model (TEMM). However, two further episodic long-term memory models, which incorporate inspiration from the ACORNS memory architecture, were developed within the ACORNS project and are the exemplar-based model and the activation based matching system. Demuynck (2009) D4.3 - 'Report on exemplar-based and activation based matching' describes these second two episodic long-term memory models in a great deal of detail. It is possible in D4.3 to find a

description of how the exemplar-based model and the activation based matching system relate to the ACORNS memory architecture and a comparison with other ACORNS memory models.

The ACORNS memory architecture inspired models perform various activities required for an automatic speech recognition system including selective attentional focus modelling, speech/non-speech attention, self-organised association of speech and semantic (visual) features, keyword recognition, early vocabulary acquisition, phone/word labelling, and building word-like units from cross-modal stimuli. These models include approaches that are new to the field of speech recognition and are more long-term with regards their likely impact on speech recognition technology. Some of the ACORNS memory models make use of speech data less complex than the ACORNS database recordings as an initial test of the approach. However, they offer the opportunity to use the ACORNS speech data in the future. This report outlines extension of the current memory models described in WP3.2 as well as new ones. Table 1 gives an initial comparison of the models produced within the ACORNS memory architecture with greater details provided throughout this report and summarised in Table 4. Table 1 gives an initial description of the memory models by indicating their applications, the memory structures from the ACORNS memory architecture the models are based on, whether the models provide results or are at the conceptual stage, and the auditory database they use. For instance, the attention gating (AG) model performs selective attention by speech/non-speech classification, takes from the memory architecture the interaction between working memory and semantic long-term memory, gives classification results, and uses the English ACORNS database recorded in period 1 and recordings of non-speech data.

The structure of the remainder of this report is as follows: Chapter 2 describes the ACORNS memory architecture. Chapter 3 is a description of attention in the ACORNS memory architecture. Chapter 4 provides an examination of models inspired by the ACORNS memory architecture's interaction between working memory and semantic long-term memory. Chapter 5 considers models based on the interaction between working memory and episodic long-term memory. Chapters 3, 4 and 5 also include comparisons between the memory models and approaches typically used in speech recognition. Chapter 6 provides a comparison of the various semantic long-term memory and episodic long-term memory models developed within the memory architecture. Chapter 7 examines the relationship between the ACORNS memory architecture and sensorimotor representations. Chapter 8 gives a summary of the findings from the rest of the report and final concluding remarks.

Table 1 Overview of the speech acquisition and recognition models developed within the ACORNS memory architecture. (WM - working memory, LTM - long-term memory, P -yes and O - no)

Model	Application	Interaction				Produce Results	Database
		WM	Episodic LTM	Semantic LTM	Motor LTM		
Special attentional focus	Attention	P	O	P	O	P	Finnish ACORNS Period 1
Attention gating (AG) model	Selected attention: Classifier speech/non-speech	P	O	P	O	P	English ACORNS Period 1 Non-speech sounds
Hierarchical Recurrent self-organising map (H-RSOM)	Speech and semantic feature association	P	O	P	O	P	English ACORNS Period 1
Restricted Boltzmann Machine (RBM)	Spoken word recognition	P	O	P	O	P	TiDigits
Hierarchical Non-negative matrix factorisation (NMF)	Early vocabulary acquisition	P	O	P	O	Conceptual stage	O
Recurrent self-organising model combined with Helmholtz machine (RSOM-HM)	Keyword recognition	P	O	P	O	Preliminary limited	English ACORNS Period 1
Acoustic DP-Ngrams	Keyword learning	P	P	O	O	P	English ACORNS Period 2
TEMM (Temporal Episodic Memory Model)	Character recognition	P	P	O	O	P	TI-ALPHA
Sensorimotor control	Sensorimotor representation	P	O	P	P	Conceptual stage	O
Exemplar-based model (D4.3)	Phone/word labelling	P	P	O	O	P	Dutch ACORNS Period 2
Activation based matching (D4.3)	Phone/word labelling and segmentation	P	P	O	O	P	Dutch ACORNS Period 2 TIMIT

2 Overview of ACORNS memory architecture

The ACORNS memory architecture (Figure 1) [ACORNS 2008] introduces speech into the Echoic memory and visual samples into Iconic memory as separate modalities. Next, the architecture introduces into working memory an attention-gated version of the current audio and visual (semantic feature) input (from Echoic and Iconic memory) that produces an activation representation of the input through weight-like structures stored in long-term memory. Semantic features are used to approximate the visual inputs of an infant learner. Learning of weights occurs based on the activations produced in the working memory, so new examples of audio and visual samples become incorporated into long-term memory. Attention mechanisms control the updating of the learned long-term memory weights for other automatic speech recognition applications. Whilst iconic memory through semantic features represents the complete scene, working memory represents, maintains and processes only the relevant object. Changing the weight-structures of the model produces and updates episodic and semantic long-term memories.

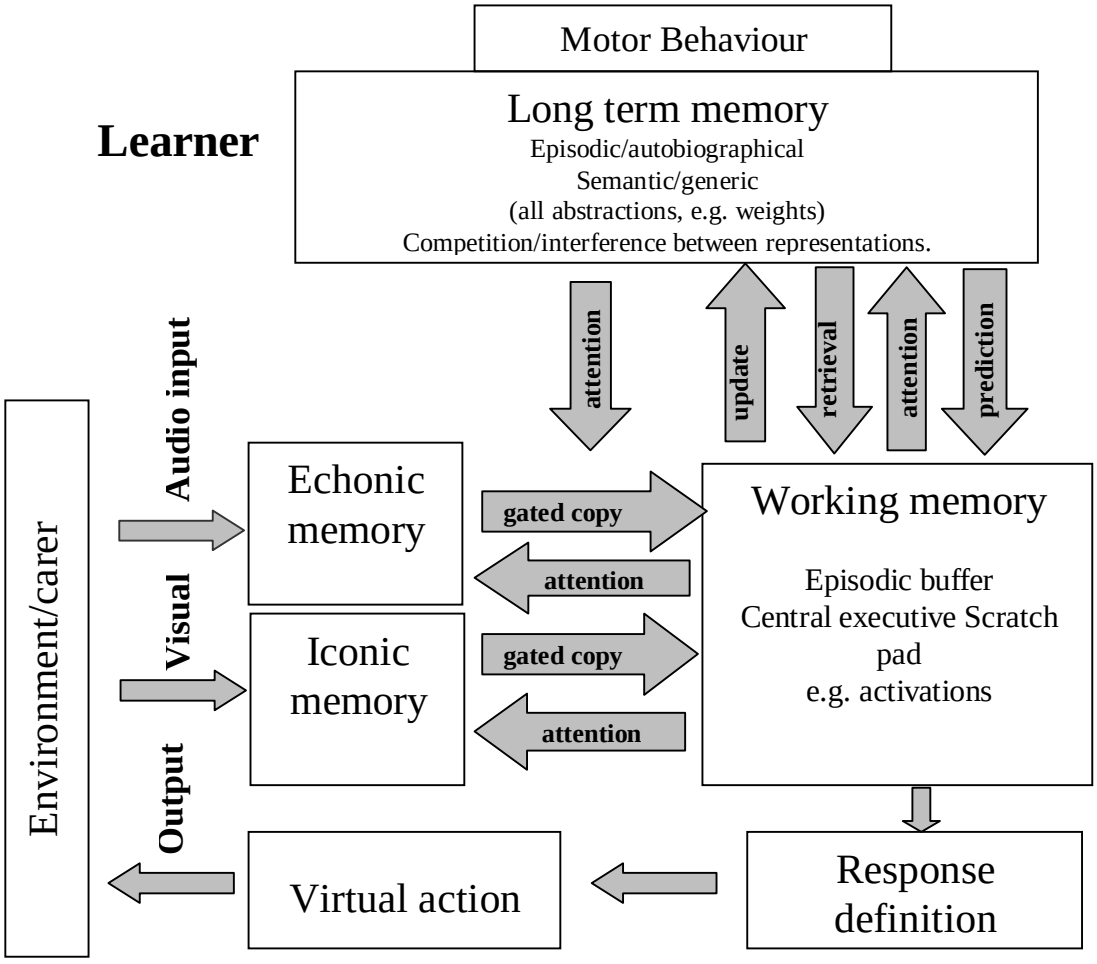

Figure 1 The ACORN memory architecture. Boxes and arrows refer to data structures and processes, respectively.

Figure 2 shows a representation of the hierarchical structure of the ACORNS memory architecture. This memory architecture offers a hierarchical organisation that allows the development of representations at speech sounds lower than the word, at the word level and finally at the utterance level. At the lowest level, the model combines the audio and semantic (visual) feature inputs with the learned weights (long-term memory) to produce representations of speech. Using the long-term memory weights W_1 and a speech input produces, in working memory, activation representations A_1 of speech sounds lower than the word. Using learned weights W_2, this provides representations at the word level A_2 by combining semantic (visual) features of words A_s with the phone representations previously produced by A_1. The semantics (visual input) activations (A_s, see Figure 2) come from the representation in working memory of semantic features. By using learned weights W_3 and activations A_2 for words this develops activation patterns for utterances A_3.

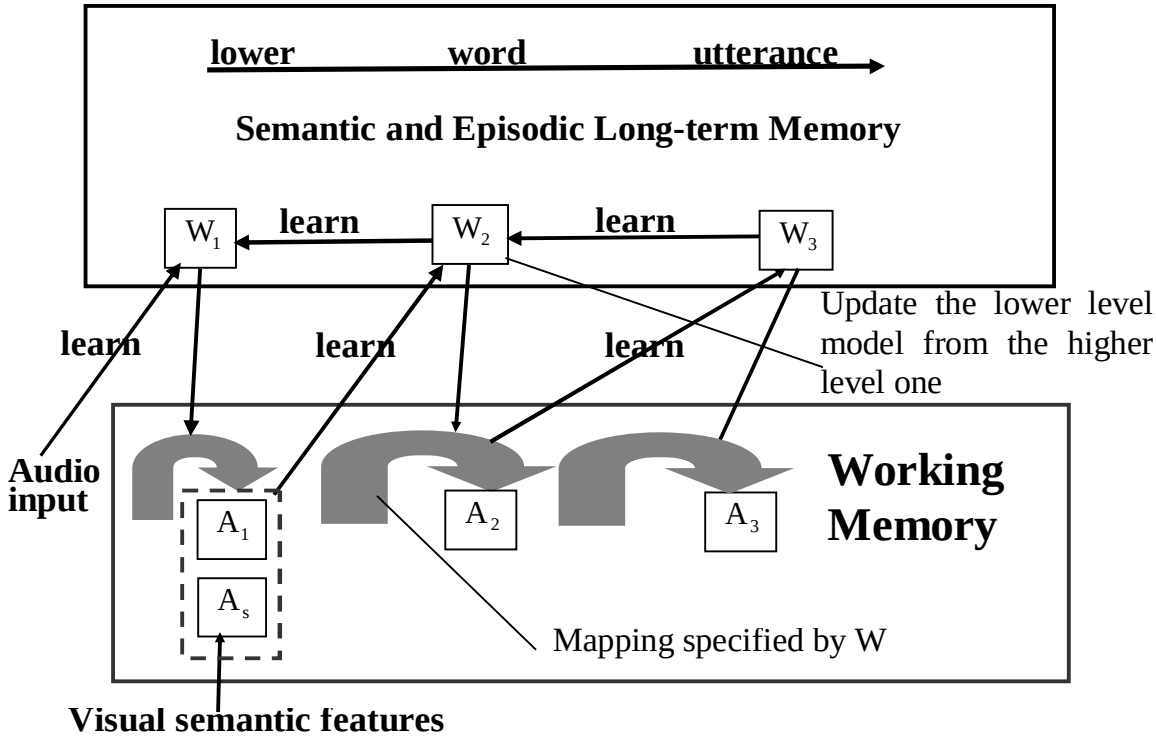

Figure 2 Hierarchical organisation of the ACORNS memory architecture. (After ACORNS 2008)

3 The role of attention in the ACORNS memory architecture

Two mechanisms were developed within the ACORNS memory architecture that focus on selective attention.

3.1 Special attentional focus model

The first memory model that provides an attention mechanism to the ACORN memory architecture provides special attentional focus to keywords in utterances to facilitate word learning and recognition. The word-learning algorithm uses a modified concept matrix approach (weight-like structures that are stored in semantic long-term memory) to track transitional probabilities of vector quantised speech, quantisation being provided by the clustering algorithm devised in WP2. Two basic attentional learning situations implemented in the current version of the algorithm are tested. In *the first situation,* the learner has absolutely no feedback from the external world except for input consisting of spoken utterances, corresponding tags, and temporal locations of the keywords. This simulates a situation where the learner gets accurate information about the keyword location from some other process, e.g., by processing of the prosody of the input. In *the second situation*, which is a so-called reinforced learning environment, the learning agent obtains feedback for its decisions from the caretaker. It was found that focused attention does not directly lead to better recognition results in this type of a learning problem, but it may help word segmentation and therefore acquisition of word models. However, a reinforced learning algorithm can also detect keyword locations with a moderate accuracy. A more detailed description of this selective attention model can be found in WP3.2 report.

3.2 Reinforcement Attention Gating (AG) mechanism to classify speech from non-speech

The second attention model developed within the ACORNS memory architecture is the reinforcement attention gating (AG) mechanism to classify speech from non-speech. This attention mechanism was original described in WP3.2 report and Elshaw et al. 2009a. The attention-gating (AG) mechanism uses actor-critic learning to perform selective attention towards speech. Through this selective attention approach, the AG mechanism controls access to working memory processing by only passing speech into the model. There has been an extension of the model to include a simple voting system.

A simple voting type system was incorporated in the AG mechanism and found to have a small positive impact on the classification rates previously achieved. The simple voting approach involves taking a window over 5 classifications of the input sections and adjusting the classification based on the majority rate. In Figure 3 the AG mechanism classifies 3 of the first 5 input sections as speech (black), and so the simple voting approach makes the first 5 input sections as speech. Although this basic voting approach did not have a major impact, this method gives a small growth in the AG mechanism's

classification rates. Using the basic voting system achieves a 1.2% improvement to 81.2% on the 'New utterances by training female speaker' data. For the 'Unheard versions of utterances used in training from training female speaker' and 'Unheard non-speech samples from training crowd scenes' data a classification improvement of 1% occurs, which takes the former's classification rate to 81% and the latter's to 94%. The simple voting system achieves an improvement to the original AG mechanism classification rate of 0.5% for the 'Utterances in the training set spoken by the second female speaker' and the 'New non-speech samples for scenarios not used in training' data.

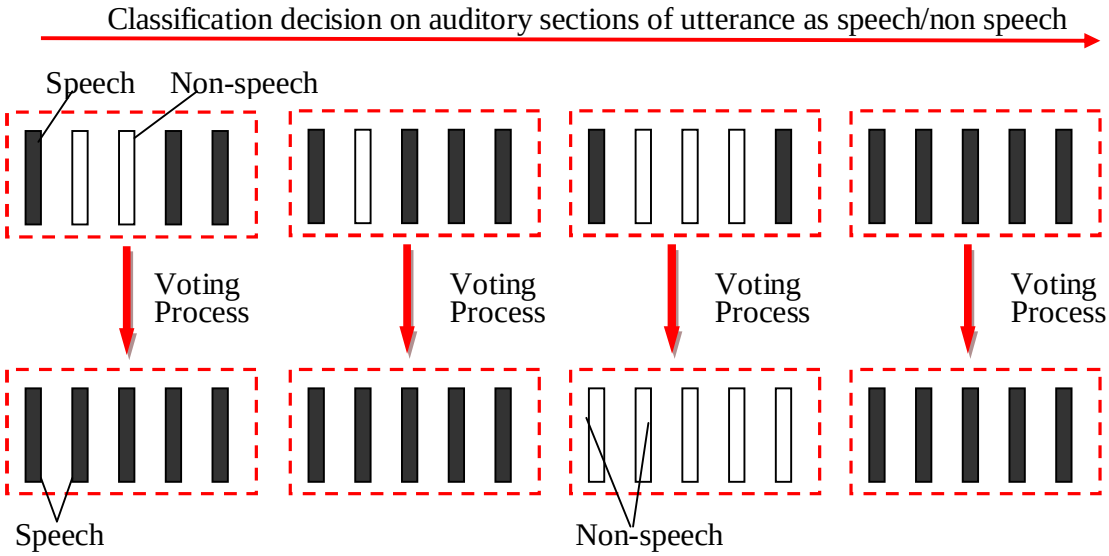

Figure 3 The process involved in the basic voting system for the AG mechanism.

As the AG mechanism uses reinforcement dopamine-like feedback, the opportunity exists to gate the input to working memory in a manner described by the ACORNS memory architecture. In the model the reward feedback occurs in a similar manner to a child giving itself an immediate reward over the auditory input and the caregiver giving a delayed reward at the end of the auditory input. The allocation of the immediate reward by a child could depend on the additional information of seeing a human nearby moving their lips in a talking manner and whether the child can produce the heard sound using their vocal cord. Furthermore, the delayed reward could relate to feedback from the caregiver who gives a reward signal if the child attends to the speech or a punishment if they fail to show interest. The AG model recreates the finding that infants learn to perform selective auditory attention by ignoring irrelevant auditory signals and concentrate on the speech signal [Mattock, and Burnham (2006)]. The AG model when compared with cognitively inspired speech/non-speech classification system, such as Shin et al. (2000) regression tree based technique, performances marginally worse. Nevertheless, the AG mechanism use of the actor-critic model [Barton et al. 1983] has shown the benefits of using ACORNS memory architecture inspiration for attention and the opportunity for further developments.

4 Interaction between working memory and semantic long-term memory associated models from the ACORNS memory architecture

Below is an examination of how the ACORNS memory architecture interaction between semantic long-term memory and working memory influences the memory models that have been developed.

4.1 Restricted Boltzmann Machine (RBM) for spoken word recognition

One of the approaches directed by the ACORNS memory architecture's interaction between semantic long-term memory and working memory is the Restricted Boltzmann Machine (RBM) [Smolensky 1986] to perform spoken word recognition. In this approach activations are created on various levels from an input representation and are stored in the working memory component of the ACORNS memory architecture. Further, the RBM model trains weights that are stored in the semantic long-term memory component to perform learning. After random initialization of weights and biases (stored in semantic long-term memory), an input vector is applied to the input layer and the hidden activations (represented in working memory in the ACORNS memory architecture) are computed from it. Using downward activations, the inputs are reconstruction from hidden representations. After that, reconstructed hidden activations are computed from the reconstructed input. Then the algorithm learns by increasing the weights and bias by the correlation of the input and the hidden units' activation minus the correlation of reconstructed input and hidden units (see Figure 4). Once the weights (stored in semantic long-term memory) for a single RBM are trained, a number of RBMs can be stacked on top of each other to produce a deep belief network [Hinton et al. 2006]. Here, the hidden units of one RBM serve as input to the next RBM.

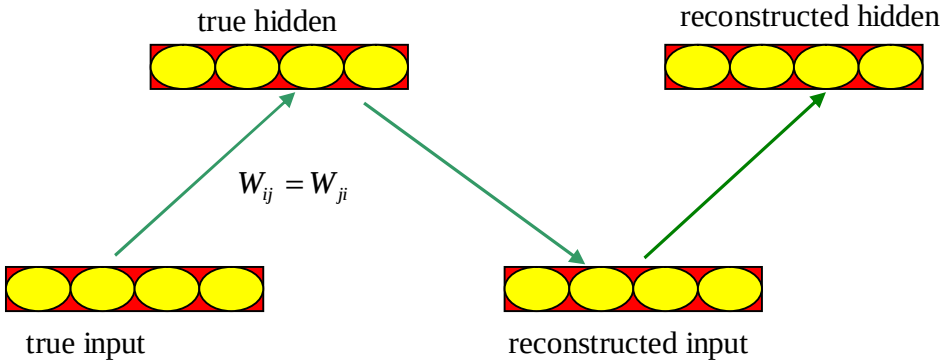

Figure 4 The RBM learning algorithm

The network is trained and tested using 27700 files from the TIDigits database of spoken English numbers. From the sound files spectral representations are computed, time-

normalised and transformed to 11 cepstral coefficients. The stimulus set is randomly split up into a training set of 26700 files and a test set of 1000 files. A three layer deep belief network is trained using the RBM algorithm. Using the hidden layer representations of the highest RBM as input, the learning using back-propagation (on both trained and untrained data) shows drastic improvement after only one training epoch; from 91.9% (chance level) to 0.66% for trained data and from 91% to 1.3% for untrained data.

This study demonstrated the ability of an approach that takes inspiration from the ACORNS memory architecture to recognise spoken digits. The performance of the model is comparable with state of the art speech recognition algorithms, but is largely unsupervised, and does not use any a priori knowledge such as phonetic labels. In contrast to other more common algorithm it creates a hierarchy of representations and it uses a biologically plausible algorithm (similar to Hebbian learning). Furthermore, we intent to replace the static representations used in this simulation with a more dynamic approach in which sound files of arbitrary length can be processes and word recognition in context will be simulated.

4.2 Hierarchical recurrent self-organising map (H-RSOM) model

A model inspired by the ACORNS memory architecture is the hierarchical recurrent self-organising map (H-RSOM) architecture for emergent temporal speech representation. This model develops a representation of speech in a temporal emergent manner by using at the lower level a speech signal RSOM and semantic feature self-organising map (SOM) and at the upper level an associator RSOM. The semantic long-term memory model approach applied to emergent speech representation uses the basic self-organising map for semantic (visual) features representation and recurrent self-organising maps [Voegtlin (2002)]. Although previously this model [described in WP3.2, Elshaw and Moore (2009), Elshaw et al. (2009b)] was trained and tested using individual words, to consider how the model performs with more realistic speech full utterances from the ACORNS database were used. The utterances are made up of carrier words and a key word. The semantic SOM now uses semantic (visual) feature representations (produced as part of the ACORNS project and described in WP5.2) for keyword nouns contained in the utterances. The H-RSOM produces activations of the current input in a model of the working memory and incorporates trained weights from various networks that are stores in semantic long-term memory.

The activations from the trained two lower components of the H-RSOM train the associator RSOM to associate the speech signal for the utterance (containing the carrier words and the keyword) and the semantic (visual) feature representations (keyword object representation). The semantic SOM input for the keyword in the utterance is introduced over the full length of the utterance as if the object is in view all the time the words are spoken. Through the association RSOM, an emergent speech representation develops that offers the possibility to associate speech and objects for grounding. Association occurs by introducing the activation values associated with each speech time slice for the utterance

from the speech signal RSOM units, the activations for associator RSOM for the previous time-step and the semantic feature SOM unit activations for the appropriate utterance object keyword. As the H-RSOM is based on the self-organisation map approach of Kohonen (1997), it is based on the best matching unit (BMU) approach that identifies the RSOM/SOM output unit that has the lowest activation value (See WP3.2 report).

Examining the BMU sub-sequences from the utterances created by the speech signal RSOM, it is possible to establish that the representations produced are associated with phone-like speech sounds. Further, the associator RSOM relates speech from utterances with semantic (visual) features for the keyword using a distinct representation based on BMU regions. Turning to the full speech signal RSOM ouput layer, the model creates distinct regional associations, based on sub-sequence of BMUs from utterances, that match phones. For instance, on an example training session units at the top left-hand area represent speech slices found in the 'T', 'CH' and 'SH' phones. Words such as 'fashion', 'shoe', 'shy', 'matches' and 'hot' include the phone speech sounds 'T', 'CH' and 'SH'.

The associator RSOM ouput layer in Figure 5 shows the location of BMU sequences related to speech sounds (from speech signal RSOM activations) for particular keywords (from the semantic feature SOM activations). The keywords (from semantic SOM) associated with a unit are represented by the colour pattern of the unit. The speech sounds (from the speech signal RSOM activations) the DARPA phonetic alphabet characters on the unit. Various units of the network associate with specific speech signal sounds and keywords. For instance, the 'D AH' speech sounds from carrier words in the utterances contain the keyword 'book' associate with unit 3 on x-axis and units 4 and 5 on the y-axis. These are the black units with a 'D AH' on them. When considering the sequences of BMUs, the 'S' speech sound from both carrier words and a keyword in utterances containing the keywords 'shoe', 'daddy' and 'book' locate close together on the map. Speech sounds (from speech sound RSOM activations) for specific keywords (from semantic feature SOM activations) locate in close units on the associator RSOM. For instance, phone-like speech sounds from words (carrier and keyword) in utterances containing a specific keyword, such as 'IY', 'S', 'IH', 'N', 'D', 'AH' and 'AY' from utterances containing the keyword 'Daddy', occur together on the associator RSOM. Speech sounds (from speech signal RSOM) for semantically related words (from semantic feature SOM) locate in near units on the associator RSOM. For instance, the phone-like sounds from the utterances containing 'human keywords', such as 'Mummy' (gray and white checkers) and 'Daddy' (grey and black strips upwards left to right), locate close together on the map.

This H-RSOM architecture develops a representation that discriminates based on phone-like sounds despite the acoustic similarity of certain phones [Kuhl (2004)]. By the RSOM model developing a representation of words in terms of phones this matches the findings of researchers in cognitive child development on infant speech encoding [Kuhl (1993)]. Kuhl (2004) notes that infants use and recognise the phonetic characteristics of speech and the retention of such speech sounds aids the extraction and development of words. In a similar manner to infants, the development of phonetic speech sound representations by

the H-RSOM architecture would aid a recognition system by providing the building bricks of speech already [Kuhl (2004)]. Although Voegtlin (2002) used this recurrent approach to represent written text inputs, this report and the WP3.2 report describes the RSOM's first use on actual speech signals. The H-RSOM approach offers an approach to ground speech in semantic (visual) features inspired by the ACORNS memory architecture.

In terms of the working memory model of Baddeley (1992), H-RSOM activation patterns recreate functionality of the phonological loop by producing representations of the current speech signal. The semantic feature SOM representation of keywords recreates part of the functioning of the visuospatial sketchpad in the working memory as it gives a representation of visual inputs. The final speech representation of the associator RSOM recreates some of the functionality of the episodic buffer, in an abstract manner, by combining of the semantic (visual) features and speech signals.

The H-RSOM despite being neural network based and taking inspiration from cognitive processing, does share certain characteristics with one of the most popular speech recognition approach the HMM. Both approaches offer a temporal representation of speech, can be phone based, and consider the likelihood that the current speech input comes from a new speech sound to change the current active state. Both approaches also face the problem that they rely on many features whose determination depends very much on previous experience and trial and error. The HMM model depends on features such as the appropriate states and transitions. A HMM makes use of probabilities that a specific outcome will follow another, which indicates in a string of speech whether one phone follows another, and so assists speech recognition in a noisy environment. Although the RSOM model does not explicitly incorporate probabilities, the including of temporal information by the RSOM ensures the learning of common sequences in speech and so a high probability, through weight values, that a sequence of input slices will follow previous slices.

Nevertheless the techniques do differ, for instance the RSOM approach does not make use of the high-level of supervised learning found in HMMs as RSOM approach offers unsupervised self-organised learning. Further, the HMM architecture acts in a 'memoryless' manner in that the conditional probability distribution for the next state given the current state does not dependent on states seen before [Karlof and Wagner 2003]. However, by feeding back the activations from the output layer RSOM previous states influence the current representation. HMMs use an acoustic model for training that typically indicate the phones in the speech and a language model that gives the words in the vocabulary and phones making up these words. Moreover, the identification of a new phone or word requires the alteration of the acoustic and language models of the HMM to incorporate this. However, the RSOM would not need any changes to produce a new representation for this newly identified phone and would represent the new word using the phone-like structures that have already developed.

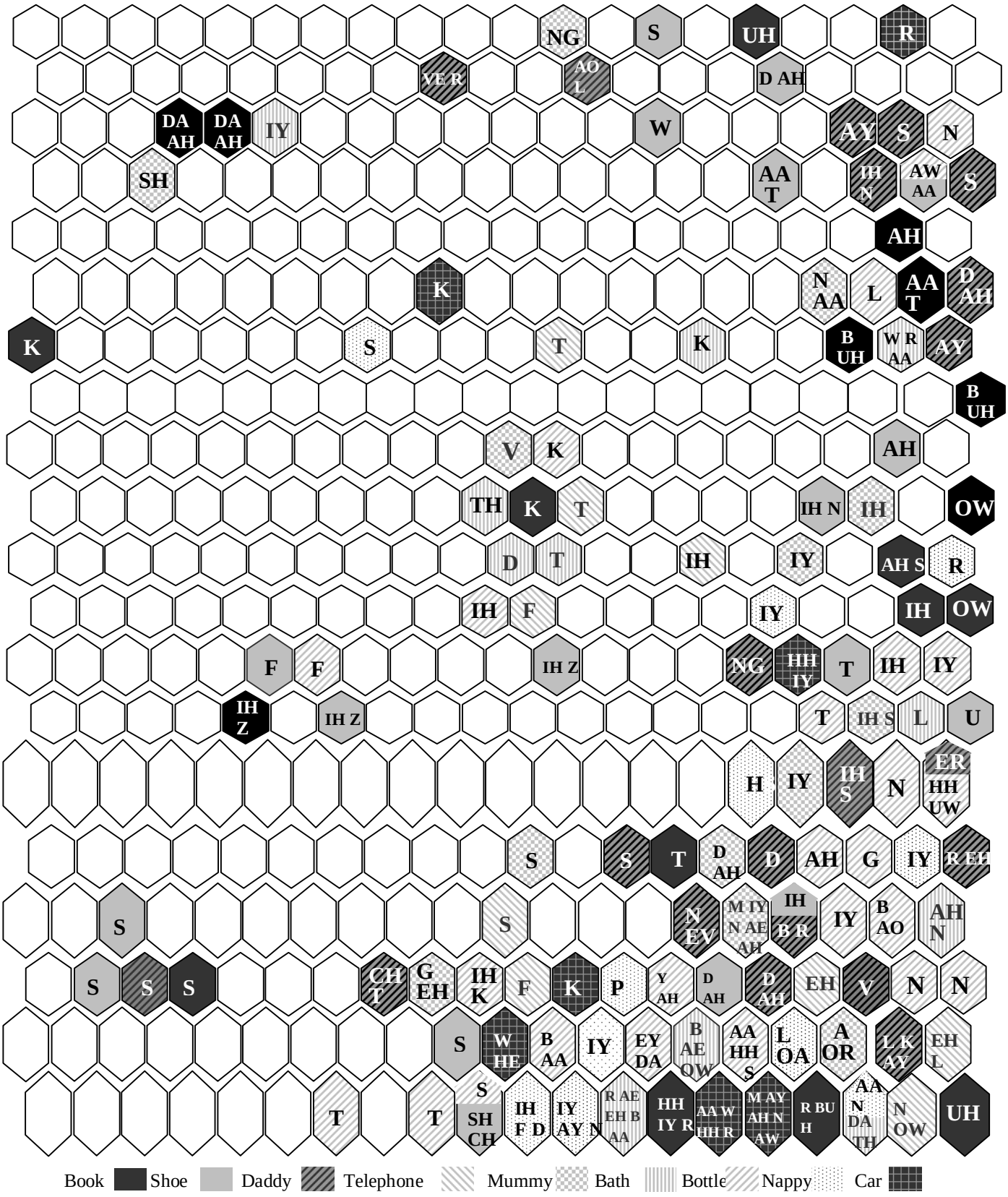

Figure 5 BMU regions of associator RSOM associated with specific phone-like speech sounds and semantic (visual) features.

4.3 Early vocabulary acquisition using hierarchical NMF

In this section of the report, we will describe how the NMF approach, which has been successfully used in the ACORNS project, can be conceptually extended by the hierarchical nature of the ACORNS memory architecture. NMF (WP4 and Lee and Seung 2001) is an algorithm in which a non-negative matrix **V** of size MxN, with M and N typically large, is factorized into non-negative matrices W and H of sizes MxR and RxN respectively, for which R<<M and R<<N. The cost function that is minimized to this end is the Kullbeck-Leibler divergence between **V** and the product **WH**. Each of the columns in **V** is closely approximated by a weighted addition of the columns in **W**, with the elements of the corresponding column in **H** serving as weights. This means that the contents of the columns in **W** can be interpreted as the models that underlie the data in **V**.

We will now describe an experiment that has been performed which demonstrates the aptitude of NMF towards the task of early vocabulary acquisition. The input data consists of Dutch sentences with a simple syntactic structure. There are four speakers (two male, two female) who provided 2000 utterances each, 1000 utterances of infant-directed speech and 1000 utterances of adult-directed speech. The result is a balanced set of 8000 utterances in total. 7000 utterances are randomly selected as the training set. The remainder makes up the test set. Each of the utterances is designed to contain a single keyword embedded in a carrier sentence.

To model the auditory periphery a spectral representation is vector-quantized i.e. mapped onto integers between one and N (the codebook size), called labels. This vectorized representation of the speech signal is called the Histogram of Acoustic Co-occurrences (HAC), and captures the spectral transition statistics of an utterance. The use of co-occurrences is psychologically motivated. It has been shown [Saffran et al. 1996] that infants during the initial stages of language acquisition are sensitive to differences in probabilities of acoustic transitions. It is clear that the HAC-representations of the words that make up these utterances can be derived from the speech utterances with NMF.

Information from several different sources can easily be combined in the NMF framework. For instance, aside from collecting static spectral information, one can also use information of spectral change by vector-quantizing *spectral velocity and acceleration* within the speech signal and accumulating the co-occurrences of those labels in a HAC-vector. Also, one is not limited to deriving co-occurrence statistics at a single time offset t. All speech information can be easily retrieved in a similar way for any number of values of t.

The experiment carried out as part of WP4 shows that NMF is well suited for learning patterns in speech. Within the hierarchical memory structure presented in this report, this opens up a number of possibilities for the NMF algorithm to act as a binding algorithm to bridge the gap from lower levels to higher levels. As an example, given an utterance, NMF could use the HAC-representation of that utterance (in this context considered to be

the lowest level information) to detect the presence of e.g. phones in that utterance. Co-occurrences of these phone-like units could then in turn be processed to detect words. We can write this as follows:

$$V_{HAC} \approx W_{HAC \to PHO} H_{PHO}$$
$$H_{PHO} \overset{co-occ}{\to} V_{PHO} \approx W_{PHO \to WORD} H_{WORD} \quad \textbf{(1)}$$

In these equations $\mathbf{H_{PHO}}$ and $\mathbf{H_{WORD}}$ are matrices that contain in each column the extent to which each speech pattern (column of **W**) is activated by the corresponding utterance, represented by columns in $\mathbf{V_{HAC}}$ and $\mathbf{V_{PHO}}$ respectively. The activations of phone-sized units, determined in every utterance, can be used to approximate the co-occurrence counts of phones in that same utterance. This way, the matrix $\mathbf{V_{PHO}}$ can thus be calculated from $\mathbf{H_{PHO}}$. Finally, $\mathbf{W_{HAC->PHO}}$ is the matrix with columns containing the HAC-representations of phone-sized units, whereas $\mathbf{W_{PHO->WORD}}$ contains word models in terms of phone co-occurrences.

Conversely, NMF could be used to learn lower-level representations from higher-level representations. For example, the HAC-representation of word-like patterns could be decomposed into HAC-representations of phone like units. Concretely, this can be written as follows:

$$W_{HAC \to WORD} \approx W_{HAC \to PHO} H_{PHO \to WORD} \quad \textbf{(2)}$$

where $\mathbf{H_{PHO->WORD}}$ indicates to what extent each phone-like unit is present in each word-like pattern. It is also possible in NMF to combine data from several levels to train speech representations, for instance:

$$\begin{bmatrix} V_{HAC} \\ V_{PHO} \end{bmatrix} \begin{bmatrix} W_{HAC \to WORD} \\ W_{PHO \to WORD} \end{bmatrix} H_{WORD} \quad \textbf{(3)}$$

In addition, meta-information similar to the semantic information in the above experiment could be added and/or combined with other data on any level of the memory architecture (we make no propositions of what this meta-information should look like, we merely wish to point out the possibility of adding it). The flexibility and simplicity of the NMF algorithm within the memory architecture provides a myriad of ways in which multiple streams of information derived from a speech signal can be combined and processed, ever refining representations of speech at all levels of abstraction within the memory. Since every part of the hierarchical algorithm has been applied with cognitive plausibility in mind, this method shows an interesting way of thinking about the deep learning processes going on within the head of infants during the early, and not so early, stages of language acquisition.

4.4 Keyword prediction combining speech signal RSOM with the Helmholtz machine (RSOM-HM)

Inspired by the ACORNS memory architecture, prelimary studies were performed to combine the speech signal RSOM from section 4.2 with a Helmholtz machine (Figure 6). For the speech signal RSOM and the Helmholtz machine model (RSOM-HM), the activation patterns produced on the two output layers relate to working memory in the ACORNS architecture and the weights relate to long-term memory. Although currently performance of this keyword prediction systems is limited (32% correct classification of keyword objects associated with the appropriate speech signal), further consideration might offer the opportunity to overcome the limitation.

The Helmholtz machine (HM) creates representations of data using an unsupervised approach. Bottom-up weights w^{bu} produce a hidden working memory representation r for an input z. Top-down weights produce an approximation of the input $\tilde{z}$ using the hidden representation. Both sets of weights (from long-term memory) are trained using the unsupervised wake-sleep algorithm [Dayan 2000]. As seen from Figure 6 the input to the Helmholtz layer represents one of the keywords, with one unit active for each keyword of the utterance. The other input is the BMU co-ordinates (x,y) from the speech signal RSOM for each speech time slice for the length of the utterance. These inputs are feed all at once into the HM hidden layer during training.

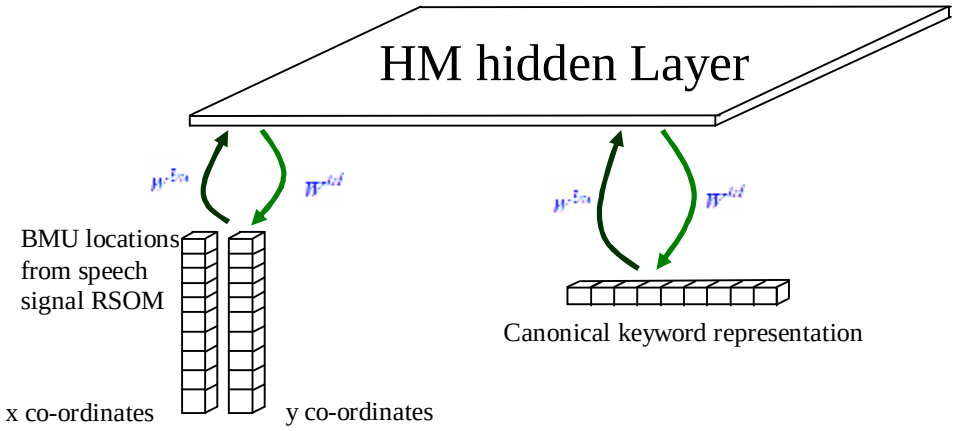

Figure 6 Training of keyword recogniser using the BMU co-ordinates from the speech signal RSOM and the keyword associated with the utterance.

The learning approach for this memory model alternates between wake- and sleep phases to train top-down and bottom-up weights. The wake phase introduces a full input z made up of the co-ordinates of the BMU from the speech signal RSOM and the canonical represention for the keyword in the utterance. The linear hidden representation

$r^l = W^{bu} z$ is obtained. The transfer function $r_j^s = e^{br_j^l} / (e^{br_j^l} + n)$, where $b = 2$ and $n = 64$ alters the linear activation into a sparse representation r^s. $\tilde{z} = W^{td} r^s$ reproduces the input and the top-down weights updated from units j to units i using:

$$\Delta w_{ij}^{td} = h r^s \cdot (z_i - \tilde{z}_i) \quad \textbf{(4)}$$

Where the learning rate $h = 0.01$.

In the sleep phase, a random code r^r initalises the activation pattern using binary activations under a Guassian envelope at a random position on the hidden layer. The input representation $z^r = W^{td} r^r$ produces a linear hidden representation using $\tilde{r}^r = W^{bu} z^r$. The transfer function above creates the sparse version $\tilde{r}^s$ from the linear representation. Updating the bottom-up weights in long-term semantic memory uses:

$$\Delta = e(w_{ji}^{bu} - z_i^r) \cdot r_j^s \quad \textbf{(5)}$$

Where the learning rate $e = 0.01$.

During training the model receives all the inputs, but during testing it ommits the canonical keyword input representation. To test the performance of the model, which combines the RSOM and the HM, it recreates the canonical keyword input for each speech slice for the utterance. The speech signal RSOM output using the weights W^{bu} produces the HM hidden layer representation (working memory activations) and from this top-down using the weights W^{td} recreates the keyword word input representation (working memory activations). Performing preliminary experiments with the RSOM-HM model indicates that the large number of different parameters involved can make the model unstable if they are not at the optimum level, and so there is a need for greater consideration of their values and how they interact for speech data. However, the model does offer an approach that makes use of inspiration of the ACORNS memory architecture and an opportunity for further developments to perform keyword recognition.

5 Interaction between working memory and episodic long-term memory model in the ACORNS memory architecture

This section allows the consideration of main models that were developed based on the ACORNS memory architecture's approach to combine episodic long-term memory and working memory.

5.1 Acoustic DP-Ngrams for keyword learning

The acoustic DP-Ngram model relates to learning keywords [Aimetti et al. 2009a] and uses techniques devised in WP2. This model takes inspiration from the interaction between working memory and episodic long-term memory in the ACORNS memory architecture to solve the language acquisition problem. The model builds internal representations by associating information from the acoustic and pseudo-visual modality. A novel dynamic programming (DP) technique segments the acoustic speech signal [Aimetti 2009, Aimetti et al. 2009b], based on repeating acoustic structure, and meaning for the keywords emerges as a property of the cross-situational statistics of the dual-modality input.

During training, LA incrementally receives utterances containing information from both modalities in parallel to build word-referent pairs. However, during testing LA only observes the acoustic modality and must reply with the concept(s) it associates with the utterance. The model learns the canonical feature representations of the concepts in the visual modality, which contains no lexical or phonetic information, to aid the segmentation process. Short-term memory, in the ACORNS memory architecture, stores both modalities of the utterance to perform *recognition* on all internal representations to discover potential 'word' candidates, and long-term memory stores these candidates as episodic memory. By using the hierarchical agglomerative clustering technique (HAC) meaning emerges through *self-organisation* as a property of the cross-situational statistics.

The DP-Ngram method discovers repeating sub-portions of two acoustic speech signals through an accumulative quality scoring mechanism. The accumulation of successive local matches yields high local quality scores that correspond with long local alignments. By backtracking from the highest local quality score this discovers the optimal local alignment. At the beginning LA has no internal representations to recognise any of the incoming utterances. When this occurs, LA stores the whole utterance as a 'word' hypothesis along with the associated visual features. Thus, LA begins life with very large and inaccurate representations before gradually sharpened with experience.

The increasing list of local alignments being stored in episodic long-term memory is denoted by $L = \{l_1,...,l_m\}$. As L continuously increases, LA clusters similar acoustic units using the HAC clustering method. The HAC method initialises each element of L as individual clusters, denoted by $\{C_1,...,C_k\}$, and then merges the two clusters C_i and C_j

with the shortest distance as defined by $d(C_{i,}C_j) = \min_{n_i \in C_i, n_j \in C_j}[d(n_i, n_j)]$, to create $k-1$ clusters. This process is repeated until $d(C_{i,}C_j)$ exceeds threshold T, creating a set of clusters each made up of local alignments with the same underlying acoustic unit. Figure 7 displays an example of the kind of acoustic clusters that may occur in long-term memory. As mentioned earlier, each element in L is also associated with any co-occurrence of the visual features that may have occurred. This is displayed in tables on the right hand side of Figure 8. Each table represents a cluster, which contains the representative local alignments $l_p,...,l_q$ along with their associated visual features, and the accumulation of the clusters visual features. C_2 builds an increasing accumulation for the visual feature '**ball**', whereas, C_1 is noisy. If we assume the acoustic representation for C_1 is [ball], then with experience LA will gain increasing confidence that it has an internal representation of a key word.

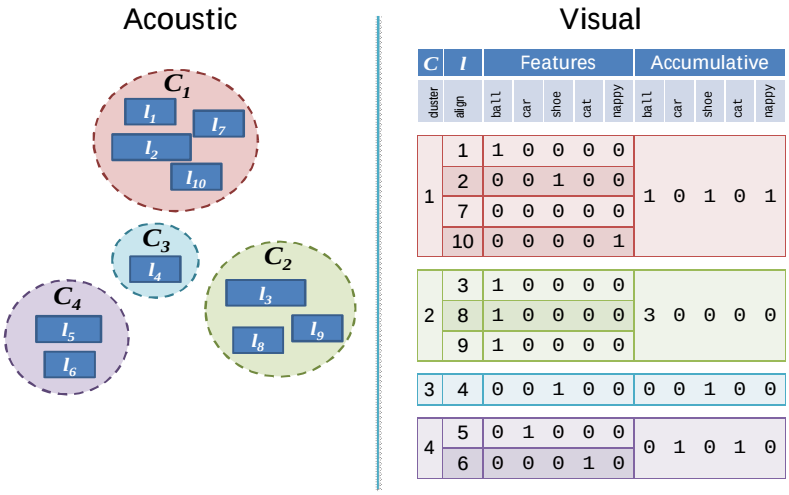

Figure 7 Diagram of the acoustic clustering of local alignments into similar word-like units and the emergence of meaning being derived from the accumulation of the semantic features within each cluster.

Cognitive science describes human development as a dynamic system, by visualising it as a continuously evolving epigenetic landscape. Figure 8 shows the epigenetic landscape for all internal representations in long-term memory. The x-axis refers to the cluster space, thus, the width of each well represents the amount of acoustic variation from the median within each cluster. Each cluster is positioned in chronological order along the x-axis, with the newest being appended to the right-hand side. The y-axis refers to the probe moment, which shows the emergence and continuous evolution of each cluster after every utterance observation (only the first 12 utterances have been drawn to preserve clarity). The z-axis refers to the semantic stability (S), which is defined as the semantic cleanliness of the cluster C_i.

After observing the first utterance LA stores it as an internal representation, which can then be used for recognition. The most common repetition is 'the', as represented by the cluster with the median token 'the'. Although there are a lot of occurrences of this item, it does not gain semantic stability. Whereas the two clusters with the median representations 'book' and 'a shoe' gradually gain semantic stability, and they represent keywords.

The design of training and test sets uses a selection of utterances recorded within the ACORNS project. The database consists of 4000 utterances spoken by two male (M1 and M2) and two female (F1 and F2) speakers (1000 utterances per speaker). The training set consists of 450 single-speaker utterances from F1, containing both acoustic and pseudo-visual information. To perform keyword detection, the model compares the test utterances with each internal representation and returns the visual features associated with the cluster that achieves the highest quality score. When considering the keyword detection accuracy as a function of the number of utterances observed the model discovers keyword representations extremely quickly but accuracy never quite reaches 100%.

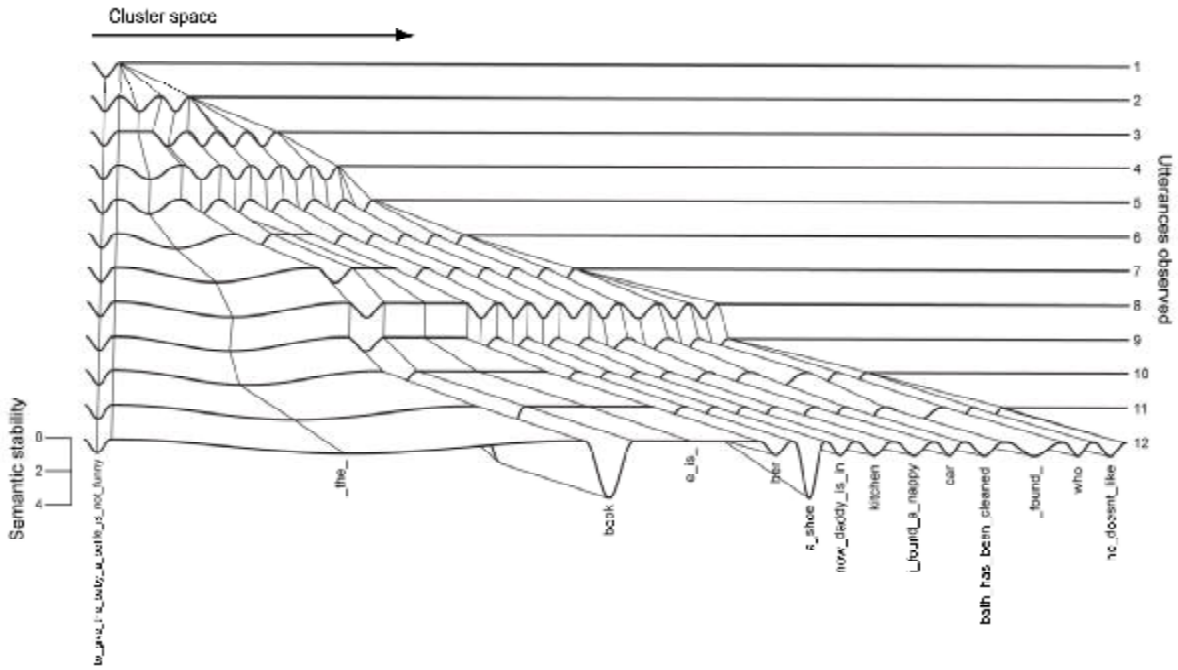

Figure 8 Epigenetic landscape of all the internal representations during the first 12 training utterances. Each cluster is displayed as an attractor well where the acoustic variation is plotted as the width within the cluster space and semantic stability is plotted as the depth. Two clusters representing an underlying keyword have already begun to emerge from the noisy clusters – 'book' and 'a shoe'.

The results for the Acoustic DP-Ngrams show that the system displays similar emergent behaviour as the dynamic systems perspective of human development [Smith and Thelen 2003] to discover keywords. Unlike the HMM, the Acoustic DP-Ngram model gains knowledge without any pre-specified linguistic rules and builds internal representations which are continuously evolving with varying stability. The results show that LA

successfully builds internal representations of keywords and can distinguish non-keyword representations by their semantic noisiness and epigenetic landscape. This information would allow us to make the system more computationally efficient by reducing the size of internal representations by getting rid or forgetting unimportant clusters.

5.2 Temporal Episodic Memory Model (TEMM)

Another episodic long-term memory model within the ACORNS memory architecture is the Temporal Episodic Memory Model (TEMM). TEMM improves on the lack of temporal sequence processing in the episodic model MINERVA2 (WP3.2). As the base operation, TEMM follows the approach found in MINERVA2 as described in WP3.2. TEMM employs a prediction mechanism as an additional source of information. The TEMM system acquires information related to how well each trace in the database fits the current input data. Feature prediction is a central part of TEMM. The fitness of the predictions to the input data and how discriminating those predictions are with respect the next best class provides an indication of (i) the goodness of the previous decision; (ii) the goodness of fit of the current data to future data. The prediction step fits neatly into the TEMM framework; by using the acquired similarity, or activation, of the traces (stored in long-term memory of ACORNS memory architecture) to input frames, it is possible to produce predictions for the features of the next input frame.

The prediction step serves as a method to generalise from training data. It also means that the model has an in-built immediate assessment of the outcome of the previous step. If the predictions fit the next input frame well, it adds credibility to the adequate use of the training data for the assessment of the input data that created the similarity/activation measures. On the other hand, if the prediction does not fit the input data at time $t+1$, then it may be necessary to question the data's assessment of classification of input frame t as well.

It should be pointed out that feature predictions can and possibly should model context-dependent predictions (depending on the data structure in the database used). For example, in a phone-based recognition task there would be merit in creating different feature predictions for different class contexts. As a consequence, the prediction step allows the model to keep track of the likelihood that the next input frame belongs to a particular class. This information matches the "intensity" of a prediction (corresponding to the summed activations that led to the prediction). I.e. a prediction's intensity corresponds to a prior expectation that the next frame belongs to the same class.

Using the concept of a "trace unit", a sequence of successive traces from the database, introduces temporal information in TEMM. The database stores traces in *sequence*. The trace that follows any one trace in the database holds the frame that followed the previous frame in the speech signal. This means that trace units are blocks of traces (i.e. frame values and class information). These trace units hold an expanding context which, due to the fact that they preserve an accurate account of sequence in the original speech signal, contains the fine temporal information. Trace units expand as a function of the confidence associated with the classification of the input frames.

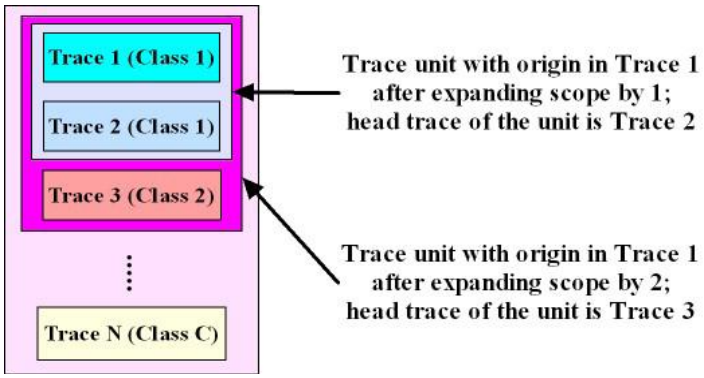

Figure 9 Schematic diagram of the process of forming trace units.

An important issue is when such trace units are formed. It was argued that the fit of the prediction to the predicted input can be interpreted not only as a confidence measure for the previous decision, but also as an indication of how well the database with the current trace activations represents the new input data. If it seems that the database with the current trace activations represent the input data sufficiently well, then it seems reasonable to allow continued use of the accumulated information. In this case, information is preserved by forming trace units that in effect are updated (trace) activations whose scope no longer only hold one single trace in the database, but a step-by-step (over time) growing sequence of frames. A second issue is how these temporal units are formed and used. With each new input frame a prediction is computed for the features in each class. In the event that the prediction matches the predicted input frame well (enough), the scope of the trace units is expanded by one frame.

By adding “transition probabilities” as an information source for updating the activations, trace units that belong to more likely classes are rewarded. This means that all trace units keep an indirect memory of how well the classes associated with its frames fit the assumed classes of the input. Class information is thus included into the trace units' activation assessment. Trace units have a scope that is not limited by class labels, they can cover one frame but, as they expand it is possible that the trace units can cover words and even sentences. How many different classes are covered in each trace unit is dependent on the underlying classes of the traces, and varies from trace unit to trace unit.

TEMM performs MINERVA2-type classification when it has no history to work from, i.e. either when the first input frame is submitted, or when a trace unit’s prediction has been disregarded as not a good fit to the data. MINERVA2-type classification means that the current class is judged on the overall activation intensity of traces belonging to each class. When, however, “history” is available, the system has much more information available, e.g. activation values for trace units which cover a larger frame context (except for the first iteration). Further, the system predicts features for each class (context) available. The predictions are computed using the trace unit’s larger context incorporating history. So it can be argued that the predictions are a class-dependent summary of all the information available to the system at the time a prediction is made. Hence, it makes sense to use such

summaries for further decisions. This means that when predictions for current input frames are available, classification decisions are based on them.

The database chosen for this investigation was the TI-ALPHA isolated word corpus because of the high confusability of the words set and its consequent high sensitivity to alternative recognition approaches. The data used consisted of two speakers, (one male (M1) and one female (F1)), uttering two letters of the orthographic alphabet –*"S"* and *"J"*. The training set consisted of 20 utterances per speaker and the test set consisted of 16 utterances per speaker. All experiments were conducted using MFCC features and a 25ms frame was taken every 10ms. Only one feature was used (for both TEMM and HMM) in order to minimize the influence of the distance measure used in TEMM. The classes corresponded to whole-word labels, and speaker-independent (SI) and speaker-dependent (SD) experiments were conducted.

TEMM is tied to one distribution estimation per feature, a single-state single Gaussian HMM was also computed for direct comparison. All HMM models were trained by incremental mixture splitting. The number of components per mixture was optimized for best performance. All references to the number of states in an HMM refer to emitting states only. In order to investigate the influence of different features on the recognition results, and their suitability for the different models, two distinct experiment conditions were set up, each using a different MFCC feature to represent the data. Condition 1 used C0 and condition 2 used C2 as features. The results are shown in Table 2 and Table 3:

Table 2 Recognition results using the C0 feature.

SD: TEMM (p=2)	33.15 %
SD: HMM 1 State (single Gaussian)	40.04 %
SD: HMM 3 States (single Gaussian)	28.81 %
SD: HMM 3 States (GMM 2)	22.99 %
SI: TEMM (p=1)	40.63 %
SI: HMM 1 State (single Gaussian)	46.38 %
SI: HMM 3 States (single Gaussian)	40.59 %
SI: HMM 3 States (GMM 120)	37.20 **%**

Table 3 Recognition results using the C2 feature.

SD: TEMM (p=2)	32.25 %
SD: HMM 1 State (single Gaussian)	32.61 %
SD: HMM 3 States (single Gaussian)	36.26%
SD: HMM 3 States (GMM 2)	36.20 %
SI: TEMM (p=1)	33.34 %
SI: HMM 1 State (single Gaussian)	72.31 %
SI: HMM 3 States (single Gaussian)	69.29 %
SI: HMM 3 States (GMM 120)	61.98 %

Comparing the performance of TEMM with the HMMs, the recognition results are rather interesting. When using the C0 feature HMMs performs significantly better than TEMM.

This may be due to the fact that C0 models overall energy in the signal. As such, neither model has much opportunity to retain fine details of the speech signal. Hence TEMM is unable to use such information to its advantage. This interpretation is supported by the recognition results using C2. Here, TEMM outperforms the 3-state HMM in the SD condition as well as in the SI condition (the difference is statistically significant only in the SI condition).

A closer analysis of the results showed that HMM output for all test conditions tended to remain in one model for a relatively long time, and hence gave rise to rather smooth recognition results that seldom changed model within a test utterance. This led to recognised words with long, (in this case) more realistic durations, and often to single-word recognition of the utterance, thereby allowing the HMM an unfair advantage to use more information on which to base its decision. This was the same even for single-Gaussian one-state HMM models. A further investigation of the model's parameters showed that the HMM's transition probabilities favoured self-transitions (i.e. transitions into the same state instead of the next), and thus the most probable cause for this output smoothing. TEMM currently has no such smoothing mechanism, and the recognised class often changed from one frame to another. This is due to the fact that the classification decision in the current TEMM architecture is based solely on the similarity of the input data to its predicted feature values. As the results of the TEMM model producing promising results on the TI-ALPHA isolated word corpus, it is anticipated that in future experiments with the ACORNS database will also produce promising outcomes.

5.3 Exemplar-based model and activation based matching system

The final episodic based models developed based on the ACORNS memory architecture are the exemplar-based model and the activation based matching system. D4.3, Chapter 5 contains a description on how these models relate to the ACORNS memory architecture.

6 ACORNS semantic long-term memory and episodic long-term memory model comparison

This chapter compares the ACORNS memory architecture inspired semantic long-term memory and episodic long-term memory models that were examined in this report. Please note to prevent repetition in deliverables, see D4.3 'Report on exemplar-based and activation based matching', Chapter 1 'Situating this Work' and Chapter 2 'Time synchronous exemplar-based matching' for a comparison between the memory models described in this report and the exemplar-based model and the activation based matching system.

The hierarchical NMF, RBM and H-RSOM incorporate features of the memory prediction framework [Hawkins and Blakeslee 2006] as they use the same network structure in all domains (auditory or visual), and perform the same basic operations in all processing domains and on all representational levels. All of the memory models examined in detail within this report except the AG model offer the hierarchical structure incorporated in the ACORNS memory architecture. However, this is achieved in very different manners for the diverse approaches. For instances, the H-RSOM achieves a hierarchical structure through the associating at a higher level the activations from self-organised representations of the speech signal and semantic (visual) features. The RBM model stacks RBMs on top of each other, and the RSOM-HM model has the RSOM network at a lower level feeding a speech signal representation into the Helmholtz machine (HM) at a higher level. The episodic long-term memory approach of the Acoustic DP-Ngram at the lower level of the hierarchy finds speech segments and at the higher level associates these speech segments with visual features. The hierarchical nature of the TEMM comes from at the lower level of the hierarchy the acquiring of information related to determining how well each stored trace matches the current input data and at the upper level the prediction mechanism.

The RBM and RSOM-HM model are similar in that they both aim at predicting canonical input features from learned weights in a top-down and bottom-up manner. A feature of the H-RSOM model that differentiates it from other ACORNS memory models is its use of semantic (visual) features in its representation as opposed to canonical features. The RBM, special attentional focus model, Acoustic DP-Ngrams and RSOM-HM model are similar as they predict the canonical input features through learning. RBM uses a static input representation of the speech; however more dynamic temporal speech inputs are used in the other memory models devised within the ACORNS memory architecture. Although a character of the TEMM episodic long-term memory model is the simple incorporate of other input modality, currently the TEMM model only includes a representation of speech in its representational process. The other models typically incorporate the two modalities of speech and vision.

When comparing the episodic long-term models and the semantic long-term models there is a clear different in the representations found in working and long-term memory. In the

episodic models there is storage of the current utterance in working memory and in long-term memory for the TEMM model there are utterances and for the acoustic DP-Ngrams previous recognised speech fragments and canonical feature representations. However, the semantic long-term memory models have more abstract representations of the inputs that combine activations in working memory and weights in long-term memory. A difference between the TEMM model and the other models is that it starts with traces already stored in long-term memory; however the elements stored in the long-memory of models such as acoustic DP-Ngrams, hierarchical NMF, RBM and H-RSOM are fully learned. The psychological and biological inspiration of the models developed within the ACORNS memory architecture is very different. H-RSOM is inspired by neurocognitive evidence on word representation (Pulvermuller 2003), semantic long-term memory, the working memory system and the unsupervised hierarchical self-organised learning found in the cerebral cortex. The AG model is inspired by the actor-critic reinforcement model, the Acoustic DP-Ngrams uses episodic long-term memory and dynamic systems theory, and TEMM is inspired by episodic long-term memory.

A common feature related to the Acoustic DP-Ngrams and hierarchical NMF is their active use of the co-occurrence of sections of speech in different utterances to identify speech components. This is not the case for other models such as H-RSOM, RBM and the speech non-speech attention mechanism which simply involve the introduction of the input one after other and the models finding structure in them. However, the episodic memory model TEMM use of prediction of the next input frame based on previous ones differentiates it from the other models such as H-RSOM, hierarchical NMF, special attentional focus attention mechanism and speech/non-speech attention mechanism whose prediction takes the form of classifying the current input frame. Unlike the H-RSOM model that segments speech into phone-like segments, the Acoustic DP-Ngrams offers automatically segments speech into word-like units and derives meaning through cross-modal association.

An important feature of the TEMM model compared with other ACORNS memory architecture models is it offers a starting point to move towards a new form of automatic speech recognition based on case-based reasoning [Maier and Moore 2009]. Much of case based reasoning research addresses solutions to very specific applications; for speech; however, such specific knowledge has been studied in great detail in the field of automatic speech recognition, but not in the field of case based reasoning. Additionally, case based reasoning, with its core principle of generalising reuse based on similarity, has strong links with minimum-distance approaches to classification. By use of a similarity comparison, case based reasoning (just as minimum-distance classifiers) sets a ‘centre of attention’. The centre of attention is therefore a fundamental, shared property between case based reasoning and exemplar-based minimum distance systems. The difference is that case based reasoning may retain more information, including procedural knowledge, on which to base its decision of centre of attention.

Centre of attention in such systems is defined globally by the chosen similarity (or activation) function as well as by the process of normalisation. Vertical centre of attention

is defined via the similarity function, and horizontal centre of attention is set via 'normalised weighting'. Normalised weighting (different to normalisation) does not mean *equal* importance of features, but instead means the importance of features based on their salience for the particular speech task at hand. This means that instead of normalising features, an automatic speech recognition system should focus on (i.e. pay attention to) features that are important for correct classification. Conventional normalisation will lead to suboptimal use of the relevant information. Figure 10 shows the resulting framework suggestion:

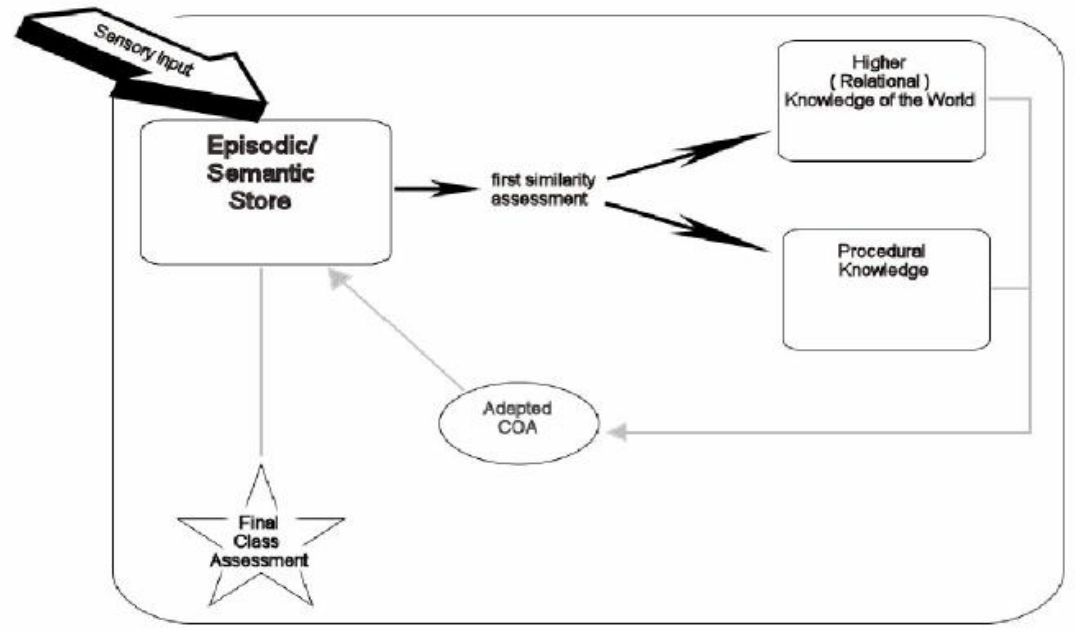

Figure 10 Proposed framework: an input triggers a first analysis of the similarity to the known traces. The found activations of the traces activate the relevant information about a) the higher knowledge of the world and b) procedural knowledge. Store experience can then help improve assessment of the data in the current step and/or future steps.

7 ACORNS memory architecture and sensorimotor representation

In order to introduce a sensorimotor module to the existing ACORNS memory architecture (Figure 1), the model has to be extended. Figure 11 shows one possible way of performing that extension. The learning agent's internal structure was divided into a sensory and a motor domain for clarity reasons. Analogous to the existing memory structure of the sensory domain, working and long-term memory in the motor domain are assumed. Similar to the long-term memory module of the sensory domain, the motor long-term memory stores articulatory gestures in the form of motor control programmes that can be retrieved by the working memory. In the working memory of the motor domain, speech production is planned and control parameters are passed from it to the speech production unit.

The memories of both domains are interconnected so that items stored in one correspond to gestures stored in the other. These interconnections could be implemented as a set of neural maps that establish mappings from one domain to the other by a Hebbian learning algorithm, a concept that was demonstrated in a computational model by Garagnani et al. (2008) after the discovery of similar structures in the human brain [Pulvermüller et al. 2009 and D'Ausilio et al. 2009]. Not only can activation patterns that emerge on the intermediate neural network layers be seen as a higher level of concept abstraction, but the across-domain activation of items can be interpreted as mirror neuron activity [Rizzolatti and Arbib (1998) and Rizzolatti and Craighero (2004)]. The phenomenon of mirror neurons would therefore emerge naturally from structures discovered in the neurological substrate.

The communication module of the previous model is replaced by a speech production unit that creates an acoustic output. This acoustic signal is directly available to the sensory input and feeds back into the speech production unit, forming an acoustic feedback loop. The error signal that governs this feedback loop is the difference between the intended acoustic signal (as represented in the sensory working memory) and the signal that was actually perceived through the sensory input. The speech production unit shown in Figure 11 is assumed to possess a means to react to this error signal in a meaningful way. This requires the learning and storage of another mapping, which was not explicitly shown in Figure 11 for clarity reasons.

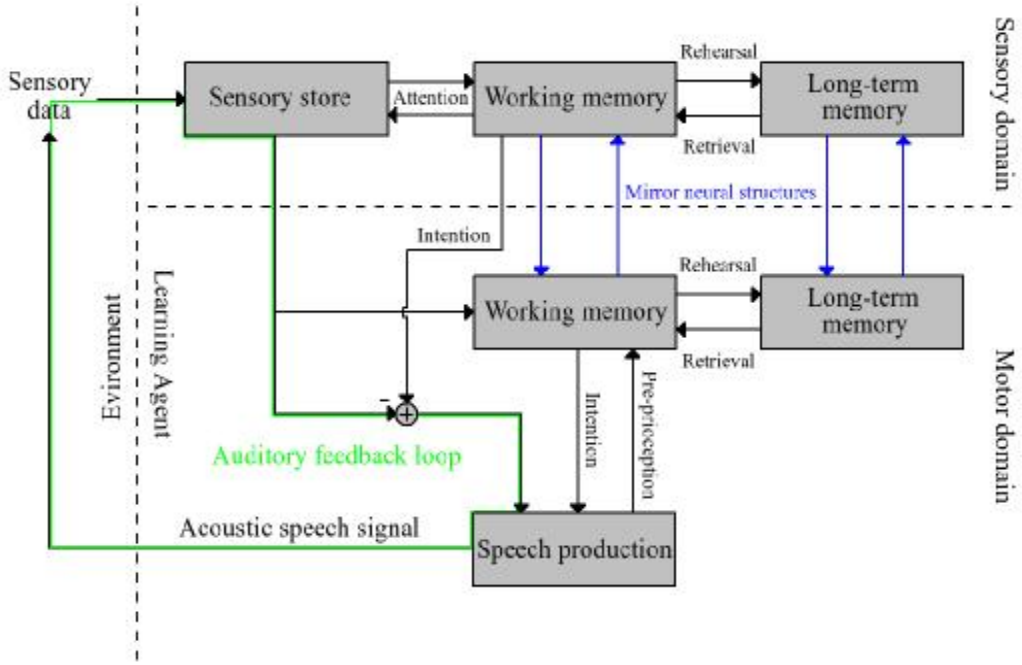

Figure 11 An extension of the existing ACORNS memory architecture to include speech production.

An important aspect of the extended model is the nature of the control signal that drives speech production. It is based on the principles of perceptual control theory (PCT, Powers 1974). PCT claims that behaviour is a reaction to a mismatch between an intention (an internal representation of what reality should be like) and reality itself (or a sensory representation of reality). In the extended sensorimotor model, two types of intentions govern speech production: an acoustic intention, formed in the working memory of the sensory domain, and a gestural intention, formed in the working memory of the motor domain. Both intentions form feedback loops in conjunction with feedback information about reality: an auditory feedback loop (marked green in Figure 11), and a pre-prioceptive feedback loop that compares real articulatory configurations with the intended ones.

The extended sensorimotor model can explain a range of common speech behaviours. For example, the model architecture contains an auditory feedback path, a structure that is assumed to play an important role in speech production generally and language learning specifically [Borden 1979]. This view is supported by speech behaviours such as the Lombard reflex [Lombard 1911]. The feedback path results almost automatically from adding a sensorimotor extension to the existing ACORNS memory model, because important parts are already present: the ability to receive acoustic input and the ability to form meaningful acoustic patterns in the working memory that can be used as a reference to form the error signal in a feedback loop. In this way, parts of the model serve multiple functions, which is a viable assumption in biological systems.

Another phenomenon that the model could help to explain is parentese, or infant-directed speech (IDS). IDS is usually higher in pitch than adult-directed speech [Kitamura et al. 2001]. In the sensorimotor model, this would mean that the representations of meaningful items in the sensory working memory would be in a form that is more likely to be reproducible by an infant's vocal tract, thus reducing the auditory feedback error signal. It could be hypothesised that normal adult speech would result in sensory representations that cannot be matched close enough by a child's speech production system in order to be accepted by the child as a acceptable imitation. Parentese would therefore bootstrap the language acquisition process on the production side.

8 Summary and final concluding remarks

Table 4 and Figure 12 summarise the findings on the ACORNS memory architecture inspired memory models described in this report. Table 4 gives an indication of how abstract the working memory and long-term memory representations are in the models; the form the speech, visual and motor inputs take; the type of prediction the models make; and whether they incorporate the hierarchical nature of the ACORNS memory architecture. Although the models are developed within this memory architecture and do share certain similarities, there are also differences between them.

Figure 12 indicates how the memory models fit into the ACORNS memory architecture and incorporate different representations into working memory and long-term memory. As can be seen, the semantic long-term memory stores the H-RSOM weights to associate speech and semantic (visual) features, the weights to classify between speech and non-speech, the RSOM-HM weights to perform keyword recognitions, the RBM weights for word recognition, for the hierarchical NMF the WH matrix for word recognition and for the special attentional focus model the modified concept matrix. The episodic long-term memory for the Acoustic DP-Ngrams model stores segments of utterances and internal representations of the utterance/canonical feature relations in the forms of clusters. TEMM episodic long-term memory stores traces as examples of speech signals. Further, this memory structure holds the speech templates from training data for the -based model, and it holds the phone and word segments from training data and K-nearest neighbour clusters for the activation based matching system.

Working memory contains the critic unit and actor unit values for the current auditory sample to perform for the AG model for selective attention between speech from non speech, and it includes the transitional probabilities for the current utterance within the special attentional focus model. This memory structure also stores the RBM model layer activations, for the hierarchical NMF the WH matrix for the test word, and the hidden layer and keyword region activations for the RSOM-HM for keyword recognition. With reference to the two episodic memory models, working memory stores for the Acoustic DP-Ngrams a representation of the current utterance and canonical features, and it stores for TEMM the prob of the current utterance and predictions as to which trace it is closest to. Working memory stores for the exemplar-based model the current input speech frame and for the activation based matching the current speech frame and the K-nearest neighbour cluster associated with current frame.

As Figure 11 above is given and to prevent Figure 12 becoming unclear, the sensorimotor control model is not shown on Figure 12. In working memory the sensorimotor control model stores activation patterns for the current speech input and motor control patterns association with that speech. The sensorimotor control model stores in motor long-term memory weights to create the motor control programmes and in semantic long-term

memory weights for the speech representation and the association between these two modalities.

WP3 successfully developed a new memory architecture that offers a framework for attention mechanism and the interaction between working memory and long-term memory. The ACORNS memory architecture is a stand alone framework and so could be used by other researchers in diverse fields to provide a supporting memory framework for their activities. The various memory models that were developed using inspiration from the ACORNS memory architecture have characters that make them different from each other. These differences show the flexibility of the ACORNS memory architecture to ensure that the models are not too restricted. The ACORNS architecture has provided a framework and direction to the various novel speech recognition memory based approach, which are real alternatives to the current state of the art. Whilst the different ACORNS models in isolation focus on different activities and working memory interaction with semantic or episodic long-term memory, within the overall memory architecture they complement each other. Further, they offer a more self-organised and emergent representation than is found in the current state of the art techniques in speech recognition such as HMMs.

Table 4 Summary of findings on models developed in the ACORNS memory architecture. P (yes) and O (no).

Model	Application	Representations Working Memory	Representations Long-Term Memory	Speech input	Visual input	Motor input	Prediction by model	Hierarchical nature
Special attentional focus model (From techniques used in WP2)	Attention	**Abstract** Modified concept matrix	**Abstract** Transitional probabilities and clusters	**Temporal** Speech	Canonical features	**O**	Class current input	**P**
AG model (WP3)	Speech/non-speech classification	**Abstract** Critic/actor activations	**Abstract** Critic/actor weights	**Temporal** Speech/non-speech	**O**	**O**	Class current input	**O**
RBM (WP3)	Spoken word recognition	**Abstract** RBM layer activations	**Abstract** RBM layer weights (bottom-up and top-down)	**Static** word representation	Canonical features	**O**	Current word representation	**P**
H-RSOM (WP3)	Speech representation	**Abstract** RSOM and SOM activations	**Abstract** RSOM and SOM weights	**Temporal** Speech	Semantic features	**O**	Current association speech/semantic features	**P**
Hierarchical NMF (From techniques used in WP4)	Early vocabulary acquisition	**Abstract** WH matrix for current utterance	**Abstract** WH matrix for previous utterances	**Temporal** Speech	Any visual representation	**O**	Current input	**O**
RSOM-HM (WP3)	Keyword recognition	**Abstract** RSOM and HM layer activations	**Abstract** RSOM weights HM top-down and bottom-up weights	**Temporal** Speech	Canonical features	**O**	Current canonical word feature	**P**
Acoustic DP-Ngrams (WP2)	Keyword learning	**Less abstract** Current utterance and canonical features	**Less abstract** Discovered speech fragments **Abstract** Internal representations as clusters	**Temporal** Speech	Canonical features	**O**	Current canonical word feature	**P**
TEMM (WP3)	Character learning	**Less abstract** Current utterance	**Less abstract** Utterances	**Temporal** speech	**O**	**O**	Character class	**P**

Table 4 (Continued) Summary of findings on models developed in the ACORNS memory architecture. P (yes) and O (no).

Model	Application	Representations Working Memory	Representations Long-Term Memory	Speech input	Visual input	Motor input	Prediction by model	Hierarchical nature
Exemplar-based model	Phone/word labeling	**Less abstract** Speech frame for current sentence	**Less abstract** Speech templates from training data	**Temporal** speech	**O**	**O**	Phone and words	**P**
Activation based matching system	Phone/word labeling and segmentation	**Less abstract** Current speech frame **Abstract** K-nearest neighbour cluster associated with current frame	**Less abstract** Phone and word segments from training **Abstract** K-nearest neighbour clusters	**Temporal** speech	Gender of speaker	**O**	Phone and words	**P**
Sensorimotor control (WP3)	Sensorimotor representation	**Abstract** Activations on motor and auditory layers	**Abstract** Weights for motor and auditory representation	**Temporal** speech	**O**	**P**	Current Motor patterns and speech	**P**

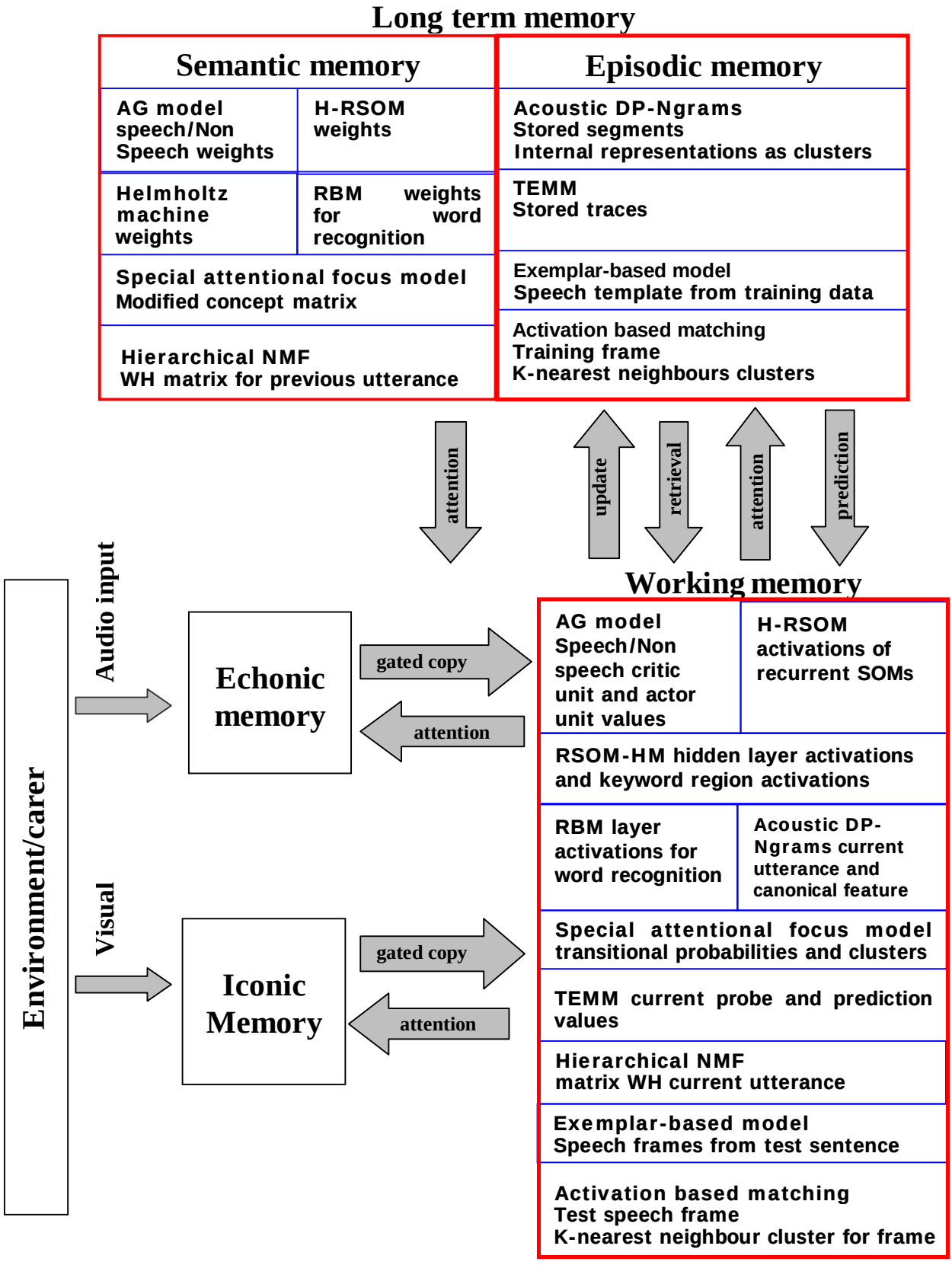

Figure 12 A representation of how the models described in WP3.3 fit into the ACORN memory architecture.

9 References

ACORNS (2008) Acquisition of COmmunication and RecogNition Skills: An overview; results of the first two years, http://lands.let.ru.nl/acorns/documents/publications/ACORNS_Summary.pdf.

Aimetti, G. A. (2009) Modelling early language acquisition skills: Towards a general statistical learning mechanism. Proc. of the Student Research Workshop at EACL 2009, pp. 1-9.

Aimetti, *G.*, ten Bosch, *L.* and Moore, R.K. (2009b) The emergence of words: Modelling early language acquisition with a dynamic systems perspective. Proc. EpiRob-09, Venice, Italy.

Baddeley, A. D. 1992 Working Memory. Science 255(5044): 556-559.

Barto, A., Sutton, R. and Anderson, C. (1983) Neuron-like elements that can solve difficult learning control problems. IEEE Trans. on Systems, Man and Cybernetics, 13, pp. 835–846.

Borden, G. J. (1979) An interpretation of Research on Feedback Interruption in Speech. Brain and Language, 7, pp. 307-19.

Burgess, N. and Hitch, G. (2005) Computational models of working memory: putting long-term memory into context. TRENDS in Cognitive Sciences, 9(11), pp. 536-541.

Davis, S.B. and Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on ASSP 28, pp. 357-366.

Dayan, P. (2000) Helmholtz machine and the wake-sleep learning. Arbib, M. (Ed) Handbook of brain theory and neural network. MIT Press, Cambridge, MA.

Demuynck, K. (2009) Report on exemplar-based and activation based matching. ACORNS Project (D4.3).

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C. and Fadiga, L. (2009) The motor somatotopy of speech production. Current Biology 19, pp. 381-5.

Elman, J. (1990) Finding structure in time. Cognitive Science, 14, pp. 179-211.

Elshaw, M. and Moore, R.K. (2009) A recurrent working memory architecture for emergent speech representation. The Bernstein Conference on Computational Neuroscience (BCCN), 2009.

Elshaw, M., Moore, R. K. and Klein, M. (2009a) An attention-gating recurrent working memory architecture for emergent speech representation. Journal of Connection Science. (To appear)

Elshaw, M., Moore, R.K. and Klein, M. (2009b) Hierarchical recurrent self-organising memory (H-RSOM) architecture for an emergent speech representation towards robot grounding. Proc. Conference on Natural Computing and Intelligent Robotics.

Garagnani, M., Wennekers, T and Pulvermüller, M. (2008) A neuroanatomically grounded Hebbian-learning model of attention–language interactions in the human brain. European Journal of Neuroscience 27, pp. 492-513.

Grossberg, S. (2003) Resonant neural dynamics of speech perception. Technical Report CAS/CNS-TR-02-008.

Hawkins, J. and Blakeslee, S. (2006) On Intelligence, Times Books, New York, NY, 2004.

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18, pp. 1527-1554.

Jones, G., Gobert, F. and Pine, J. (2007) Linking working memory and long-term memory: A computational model of the learning of new words. Development Science, 10(6), pp. 853-873

Karlof, C. and Wagner, D. (2003) Hidden Markov model cryptanalysis. Tech Report No. UCB//CSD-03-1244, Computer Science Division (EECS), University of California, Berkeley, California.

Kitamura, C., Thanavishuth, C., Burnham, D. and Luksaneeyanawin, S. (2001) Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language, Infant Behavior and Development 24(4), pp. 372-392.

Kohonen, T. (1997) Self-organizing maps, Springer-Verlag, Heidelberg, Germany.

Kuhl, P. (2004) Early language acquisition: cracking the speech code. Nature Reviews Neuroscience, 5(11), pp. 831-843.

Kuhl, P. (1993) Early linguistic experience and phonetic perception: implications for theories of developmental speech perception. Journal Phonetics, 21, pp. 125-139.

Lee, D. and Seung, H. (2001) Algorithms for Non-negative Matrix Factorization. NIPS.

Lombard, E. (1911) Le sign de l'élévation de la voix, Ann. Maladies Oreille, Larynx, Nez, Pharynx, 37, pp. 101-119.

Maier, V. and Moore, R.K. (2009) The Case for Case-Based Automatic Speech Recognition. Proc. Interspeech 2009, Brighton, England.

Mattock, K. and Burnham, D. (2006), Chinese and English infants' tone perception: Evidence for perceptual reorganization. Infancy, 10(3), pp. 241–265.

Powers, W. T. (1974) Behavior: The Control of Perception. London: Wildwood House.

Pulvermüller, F. (2003) The neuroscience of language: On brain circuits of words and serial order. Cambridge, UK, Cambridge University Press.

Pulvermüller, F., Shtyrov, Y. and Hauk, O. (2009) Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. Brain and Language 110(2): 81-94.

Rizzolatti, G. and Arbib, M. A. (1998) Language within our grasp. Trends in Neuroscience, 21, pp. 188-94.

Rizzolatti, G. and Craighero, L. (2004) The mirror-neuron system. Annual Review of Neuroscience, 27, pp. 169-92.

Saffran, J., Aslin, R. and Elissa, L. and Newport, E. (1996) Statistical learning by 8-month-old infants. Science, 274(5294), pp. 1926–1928.

Shah, J., Iyer, A., Smolenski, B., and Yantorno, R. (2004) Robust voiced/unvoiced classification using novel features and Gaussian mixture model. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 17-21, Montreal, Canada.

Smith, L. B. and Thelen, E. (2003). Development as a dynamic system. Trends in Cognitive Sciences, 7(8), pp. 343-348.

Smith, L., and Yu, C. (2008) Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition, 106, pp. 1558-1568.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D. E. and McClelland, J. L., editors, Parallel Distributed Processing: Volume 1: Foundations, pages 194-281. MIT Press, Cambridge, MA.

Voegtlin, T. (2002) Recursive self-organizing maps. Neural Networks, 15(8-9), pp. 979-991.