ACORNS

Acquisition of COmmunication and RecogNition Skills

An overview; results of the first two years

All project members
Template: 18th of September
Draft: 5 December 2008
Final: 20 December 2008

Table of Contents

1. Introduction

The aim of the European Future and Emerging Technologies (FET) project ACORNS is to develop *computational models* that demonstrate the capability to acquire language and communication skills on the basis of sensory input. In a farther-reaching perspective we need these models to explain how infants can learn to communicate by means of spoken language and to build automatic systems that substantially outperform extant Automatic Speech Recognition (ASR) systems that implement some kind of pattern recognition where the patterns are pre-defined by the system developer.
In ACORNS, the primary input from which the artificial agents acquire language is restricted to speech utterances that refer to observable physical objects and events in the environment. To ground the meaning of the spoken utterances, they are accompanied by some representation of a virtual world to which they refer. In addition, the learning agent receives (auditory only) input in the form of feedback from the (simulated) care givers with whom it communicates. The learning agent in ACORNS is passive, in the sense that it has no means for producing speech-like sounds, nor is it able to actively explore the surroundings.
The five partners in the project have their main backgrounds in speech technology and ASR in particular (although their focuses may differ). But all subscribe to the idea that for ASR to reach human-like performance a completely different approach is needed, inspired by human processing. This explains why we expect that investigating whether a new approach to modeling language acquisition and processing –inspired by a theory of cognition and intelligence- will eventually open new perspectives for ASR. It also explains why many of the methods and approaches used in the project have their roots in ASR research.
Thus, the challenge of ACORNS is to model an infant who has *no prior knowledge* about grammar, words, or speech sounds when it is confronted with the first communicative utterances produced by its care givers. By doing so, we avoid the error of reference in Artificial Intelligence: modeling some meta-level description of a process rather than the process itself (Pfeifer & Scheier, 1999). Inspired by a hierarchical model describing human memory (the memory-prediction model, developed by Hawkins 2004), the hypothesis will be tested that processing multimodal input will result in the emergence of hierarchical representations of speech that may or may not reflect the units and representations that play a pivotal role in all existing theories of language structure and language processing. To the extent that units reminiscent of words and phonemes do emerge, we will investigate whether they emerge in a fixed order, or that, alternatively, different units emerge as they are needed for communicative processing of sensory stimuli.

The literature related to modeling first language acquisition is growing fast. To position the research in ACORNS we find it useful to refer to the recent overview paper by Kaplan, Oudeyer and Bergen (2008) who distinguish five major –but not necessarily orthogonal– approaches (or stances, as they call them) in the general field of language learnability. These 'stances' are the generative perspective, the statistical approach, the embodied/social perspective, the perspective from the child's development, and language evolution. The first *Generative* stance is mainly concerned with grammaticality, and is therefore outside the part of the field that ACORNS intends to cover (irrespective of the fact that this stance is losing appeal for several other reasons). Actually, the original ACORNS proposal took position against the generative stance. Since ACORNS does not address the question of *language evolution*, the fifth stance does not apply either. However, ACORNS fits in three of the five approaches, viz. statistical learning, embodied & social cognition, and developmental learning. And, of course, ACORNS is all about computational modeling as the method of choice for investigating and explaining behavioral observations about the acquisition of language and communication skills; and for deriving novel hypotheses that can be tested in new behavioral experiments.

2. Overall approach in the project

The design of the ACORNS project is such that the work is divided into three years. In the first year the infant (Little Acorns) should be able to know when s/he is addressed and learn to understand 10 'words' (i.e. utterances that refer to 10 different objects). In the second year s/he should be able to learn 50 words, and for the final year the target is 250 words.

In order to accomplish these aims, ACORNS intends to develop and record three speech databases, a different database for each year in the project, corresponding to three development stages of the learner model. These databases should have a basic ecological validity. However, ACORNS is not trying to mimic all aspects of the actual language acquisition process, if only because there is no simulation of speech production. The databases should make it possible to test the main claims of the project, and should not take too much time to make. Specifically, since Little Acorns has no innate knowledge of linguistic units such as words, syllables or sounds, detailed annotations of the utterances in the databases are not required. It is enough that we can be sure that each utterance refers to specific objects or events.

The databases form the bases for performing simulation experiments. The experiments are all meant to investigate whether an artificial agent can learn language without endowing this agent with linguistic knowledge. In doing so several different computational approaches are investigated. These approaches will be compared with a focus on what they can tell us about the processes involved in language acquisition. Such a comparison can be instructive even if different approaches yield different performance levels in terms of the number of 'words' learned, or the speed or ease with which new 'words' are learned, etc. At the end of the project we hope to able to draw conclusions related to the cognitive plausibility of the approaches to modeling first language learning that we have investigated, and suggest promising new ways for designing and building more powerful and capable ASR systems

In the remainder of this section we present the memory architecture that lies at the basis of all research in ACORNS. In section 3 we introduce the databases for the first two years. In section 4, a number of experiments related to the different approaches to modeling language acquisition are presented. Finally, some preliminary conclusions are drawn and ideas for the final year are presented.

2.1 Memory Architecture

ACORNS is about the feasibility of the memory-prediction framework (Hawkins, 2004) as a basis for understanding language acquisition and communication. The memory-prediction framework is extremely appealing, mainly because it is based on solid and neuro-physiological evidence that has been known for a long time (Mountcastle, 1978). However, at the start of the ACORNS project there was no complete computational implementation of the framework, and it was unlikely that such software would materialize any time early in the lifetime of the project (Hawkins, 2005). The software based on Hierarchical Temporal Memories (HTMs) under implementation by Numenta™ that came closest to our needs (George and Hawkins, 2005) appeared to be too limited for reaching the goals of ACORNS (van Doremalen & Boves, 2008).
While the concepts underlying the memory-prediction framework should be the guiding principle, ACORNS never wanted to commit to one single software implementation. The Technical Annex listed several different approaches to the problem of discovering structure in speech (and visual) input and building hierarchical representations. Equally importantly, it would not be appropriate to ignore the extensive literature on memory processing in psychology research (Baddeley, 1992) that does not necessarily map one-to-one to the structure suggested by the memory-prediction framework and certainly not to its implementation in terms of HTMs. Therefore, much time and effort has been spent during the first two years of the project to design a memory model that at once reflects the results of decades of psychological research and the basic tenets of the memory-prediction framework.

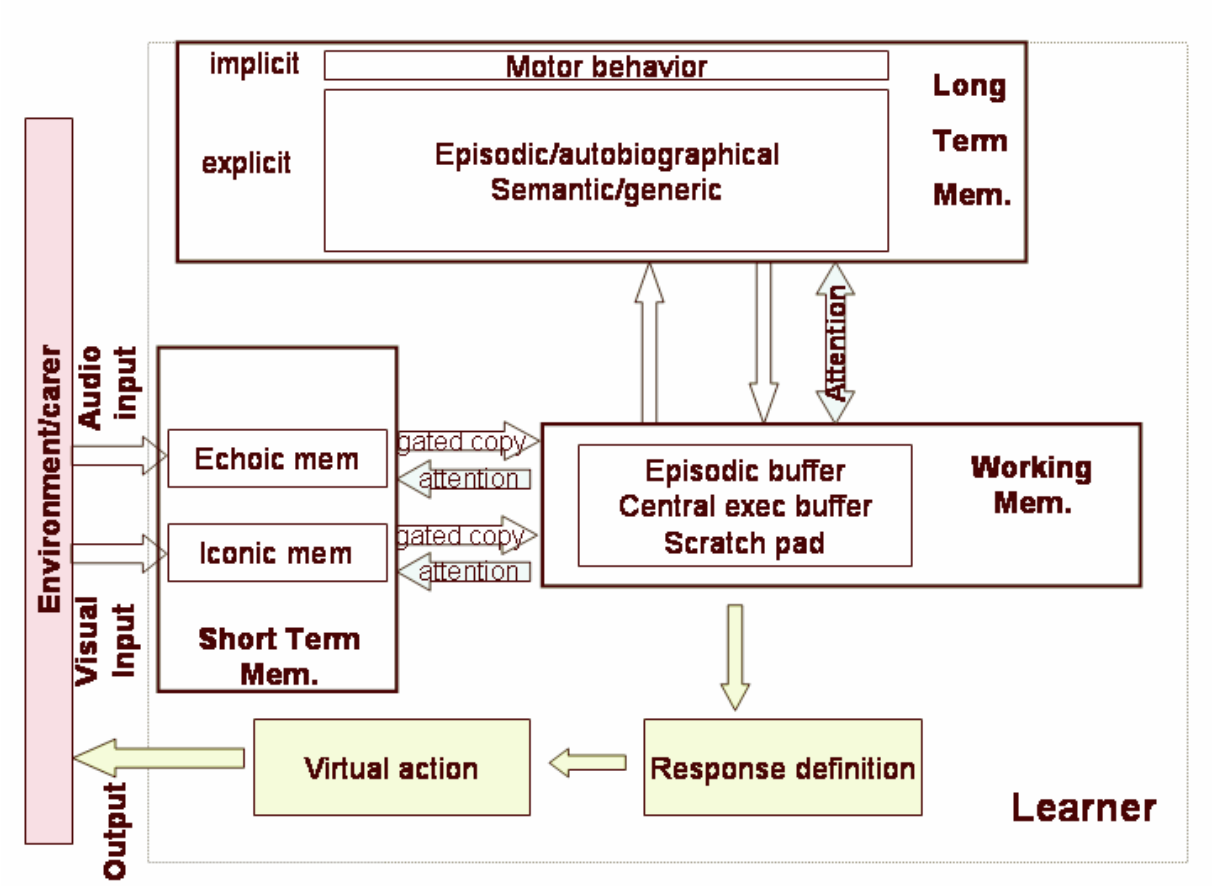

Figure 1Hierarchical modular memory and processing architecture that reflects the results from research in Psychology on Memory, Language.

The latest and most elaborate version of the model that we intend to implement is shown in Fig. 1. It reflects a widely accepted modular structure in which one can distinguish a sensory store, a working memory (or short-term memory) and a long term memory. At first sight the architecture in Fig. 1 seems to have little in common with the structures suggested by the Memory-Prediction framework, one possible representation of which is shown in Fig. 2. However, in comparing the pictures one must keep in mind that both 'models' are quite general and abstract, and that many essential details are left, perhaps as 'implementation details'.

We think that one can map the sensory store in the architecture of Fig. 1 onto the lowest level of the hierarchy in Fig. 2, if only because neither model makes hard claims with respect to the neural encoding and representation of the sensory signals at the lowest level of the cortical hierarchy. In a similar vein the processing that is going on in the working memory in Fig. 1 may very well map onto the connections that are formed and the information that flows in the higher levels of the structure depicted in Fig. 2. And when it comes to the long term memory in Fig. 1, this too must be represented in the form of connections between brain cells in the cortex. Therefore, we take it that an architecture such as depicted in Fig. 1 can implement the basic operations in a Memory-Prediction framework. Perhaps the most important difference between the models depicted in Figs. 1 and 2 is that the modularity suggested by the first model may make it easier to develop computational approaches that rely on explicit representations of speech and 'meaning' on a number of distinct levels of some hierarchy. The model of Fig. 2, on the other hand, is probably more akin to computational approaches inspired by Neural Network techniques.

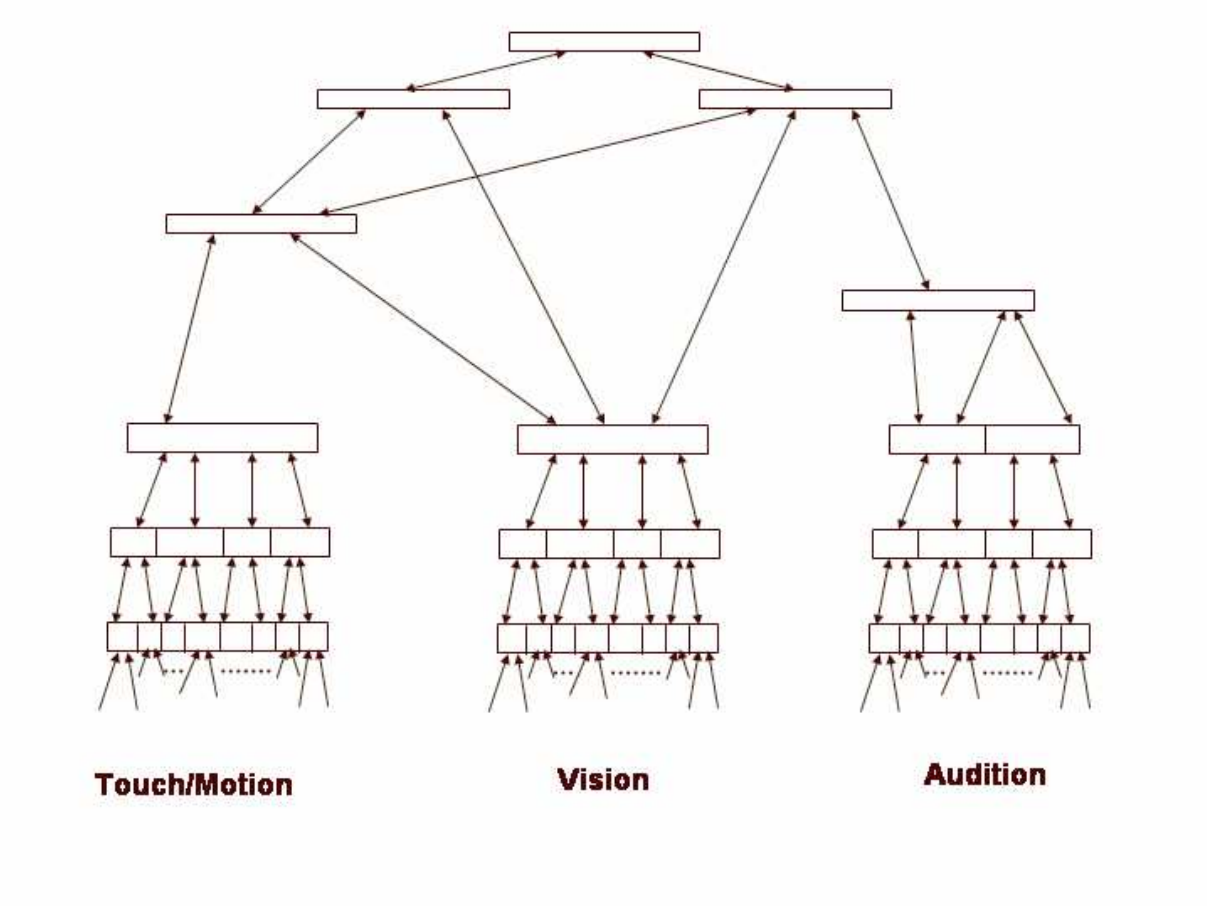

Figure 2 One possible view of the cortical hierarchy in the Memory-Prediction Framework. (After Hawkins, 2004)

3. Databases developed in the project

3.1 Year 1 database

The design of the first year's database was a topic of a lengthy discussion in the ACORNS consortium. Several ecological constraints were contrasted with various other constraints concerning e.g. acoustic variability and complexity of the carrier phrases. The eventual idea behind the first year database was to start with very simple utterances by 4 different speakers, two male and two female. The database has been recorded for Dutch, English, Swedish, and Finnish.

There are 10 target words per language. The words were selected based on language acquisition literature (i.e. 'words' reported to be known by 8 months old infants). Each of the words were used in 10 different contexts (carrier sentences), each of which was repeated 10 times by each of the speakers. Examples of sentences are "Show me the book", "Where is Daddy", where the underlined word represents the object referred to by the utterance. The target words mostly occurred in utterance final position (which is usual in infant directed speech in the languages under study). In the database, each utterance corresponds with one *tag* (i.e. code of the visual concept). The tag is 1-1 with the single keyword uttered in the utterance, and is thereby uniquely and non-probabilistically defined.

The resulting set of 1000 utterances per speaker were produced in two modes:

- a normal 'adult-addressed speech' (ADS) mode
- acted 'infant-directed speech' (IDS) mode, with exaggerated intonation and slow tempo, as if speaking to an infant about 8 to 10 months old. On advice by Elisabeth Johnson the elicitation was facilitated by the presence of a real-life picture of a young infant.

The resulting number of 2000 utterances per speaker per language proved to be a good starting point for the learning experiments.

3.2. Year 2 database

The database for the second year is meant to show that Little ACORNS is able to learn fifty "words". The target words are based on the list of "words" infants of about 12-15 months old are reported to understand. Compared to the first year database, the phonetic context of the target words is much more varied.

In contrast to the 'one-tag-per utterance' annotation of the first year database, for the Y2 database we decided to use semantic features. The major reason for this is that one of the core goals of the ACORNS project is not to use a prior knowledge about linguistic units. The tags used so far are fully deterministic and, therefore, they can be interpreted as a very strong form of a priori knowledge. During first word acquisition a child is obviously not confronted with anything resembling unambiguous meanings. Rather infants perceive objects which fall into a number of categories. For example: a red sleeping bear falls into the categories *red*, *sleeping*, *bear*, as well as possibly *animal*, *furry*, *big* or *small*, *scary etc*. And all those words can appear in a perceived utterance in the presence of the very same individual object. However, at the same time, many of the words that might also be used to denote properties of such an object might not appear. Feature-coding of objects and events allows us to replicate this situation more accurately and, in this way, eliminate a priori knowledge.

Practically, this is done by using several object position slots. Every slot can contain an individual object (including persons). An object is defined by a set of features, depending on which categories apply to it. For example one position might be filled with the feature sets of *red*, *furry*, *eats*, *bear*, and *animal*, while another slot might be filled with the feature sets of *round*, *green*, *apple*, *food*. The learning system might then be exposed to an utterance such as "The bear eats the apple." or "The red furry animal eats the round green food".

There are many possibilities of coding meaning in features. A task force involving most ACORNS partners developed a suitable coding scheme. The most appropriate way appeared to be the use of features and anti-features (i.e., *green* and *not_green*) with continuous intensity values (with an additional value for certainty), which allows us to distinguish absence of a feature from ignorance about a feature. The task force further came up with a number of (mathematical) constraints that has to hold about the feature set and the coding of objects with features (e.g. about the distance between concepts in feature space, or the relation between features and anti-features).

The words of the Y2 database were chosen with the goal of describing a simple scene with persons, objects and actions that are likely to occur in the environment of a child. For every word, a feature coding was established. As far as possible this coding was based on existing semantic feature databases. Further, the coding reflects visual semantic features, similar to those that could be extracted from visual information available to a child.

To generate the sentences of the database a scenery was used to create a list of objects (including persons), properties (colors, shapes, sizes) and actions (very few). Based on this list, sentences were created such as "there is a lion and a duck", "do you like a big cookie". In addition, specific feedback utterances were recorded, e.g. "no I mean the RED ball" for use in lightly supervised learning algorithms that require feedback for reducing classification error rates.

Next, the sequences of words in the sentences were used to create the corresponding meaning presentations. For a sentence like "Daddy sees the red ball." at least two individuals or objects (but possibly more) need to be present in the scene: one individual who needs to have the features of "daddy", but also for example the features of "man" (and in future research also "see") and another object with the features of "red" and "ball" (but also with the features of "round", "toy" etc.). Note that in the Y2 database there are more target words per utterances (up to four, average 2.8).

There are 10 speakers; 4 speakers are the same ones who recorded the year 1 database and 6 new speakers. The databases were recorded for English, Dutch, and Finnish. The 6 new speakers only produce a subset (600) of the 2000 utterances, and thereby serve as new 'previously unseen' persons. Per language, each target word occurs at least 50 times across the entire database. At the time of this

writing no results obtained with the Y2 database have been published. Therefore, this paper will focus on results obtained with the T1 database.

The database is further dedicated to experiments on the use of intonation/focus, and the interaction between the learner and care giver (especially the role of feedback in the update of internal representations). This database is not very well suited to investigate the reuse of emergent sub-word units for facilitating learning of new words. To investigate this latter issue, other existing large scale databases such as TIDIGITS, Resource Management and Wall Street Journal have been used.

4. Experiments

In this section we give a summary of the major lines of experiments that have been explored up to now. More detailed reports are available in the form of workshop and conference papers published as part of the project. All published papers can be found on the ACORNS public website[1].
All experiments refer, in some way or another, to the basic memory models shown in Figs. 1 and 2, and to the objectives for the first two years, i.e., to show that an artificial agent can build representations of speech signals without imposing a priori knowledge of meta-level concepts such as words, syllables, phonemes, etc. and that this agent can use those representations to associate novel speech input with a limited number of objects in the environment. The latter capability might be referred to in terms such as 'recognize' or 'understand', and indeed, if we define recognition or understanding as 'showing the expected response' this equivalence is fully justified.
Most of the experiments conducted so far focus on discovery of structure in what can be called a one-level hierarchy. There are two major reasons for this limitation. First, it was felt that we needed to start with an in depth investigation of the capability of the structure discovery approaches that we had available at the start of the project to find recurrent structure, without using unjustifiable pre-existing knowledge. Second, experiments with elaborate multi-level hierarchies require the availability of operational implementations of a hierarchical memory architecture. As said before, such software is not readily available. Our attempts to specify the implementation of a hierarchical memory architecture has shown that there are a large number of issues that may look at the surface as 'implementation details' but that on second thoughts appear to be serious gaps in our understanding of the structure and the operation of the brain. These difficulties not withstanding we have conducted experiments aimed at investigating multi-layered structures. Perhaps the most interesting results of these experiments so far is the finding that it may be difficult, if not impossible, to find a single processing strategy that will be optimal (or even effective) at all layers, for all types of information (or units) and for all purposes. Perhaps, this finding should not come as a complete surprise, even if one might interpret the architecture in Fig. 2 as suggesting that processing should be homogeneous from top to bottom because the structures seem so similar. It may be that differences in interconnectivity and in the type if information stored at different levels should imply different computational processes and different algorithms. In many ways the modular architecture of Fig. 1 already points in that direction.

So far, we have investigated five different approaches to the problem of discovering structure in acoustic input (that comes along with references to objects in the environment) without any form of prior segmentation (nor any form of a priori linguistic concepts to guide the process) where the continuous speech signal is represented as an acoustic waveform. These approaches exploit, in one way or another, statistical regularities that are present in the speech signal. These four research directions are Non-negative Matrix Factorization (NMF); Multigrams; DP-Ngrams and Computational Mechanics Modeling (CMM) and State Transition (or Context) Matrices. Results obtained with Context Matrices have not yet been published; therefore, we will not discuss this approach in detail in this paper.
In addition to approaches that do not start with segmentation at some linguistically motivated level, an approach based on bottom-up phonetically inspired segmentation has been explored.

[1] http://www.acorns-project.org

Finally, experiments with Artificial Neural Networks (more specifically Self-Organising Maps) have been conducted.

In their default implementations the structure discovery methods mentioned in the previous paragraph operate on large batches of input stimuli. Obviously, this is not in accordance with the way in which infants acquire language. For that reason substantial effort has been invested in attempts to modify the NMF and DP-Ngram methods to make them suitable for *incremental* learning, in both cases with some success. Making NMF and DP-Ngrams incremental makes it easier to investigate the cognitive consequences of learning in the absence of a priori defined concepts justified by meta-level descriptions.

So far, the basic representations of the acoustic signals used for most of the experiments have been conventional Mel-Frequency Cepstrum Coefficients (MFCC) in an implementation developed in WP1. During the third year of the project we will repeat crucial experiments with more advanced acoustic representations that take advantage of our knowledge about auditory processing. These features are at present under development in WP1. It will be interesting (and a task left for the third year of the project) to investigate possible relations between the auditorily salient features under development in WP1 and the results of experiments that investigated auditory processing of infants (Saffran et al., 2006).

While MFCCs have proved their usefulness in automatic speech recognition, these features are not particularly powerful representations of prosodic features. In effect, only 'loudness' has a fairly direct representation, while 'pitch' is only implicitly encoded in a manner that is difficult or impossible to decode. Yet, prosody is known as a potentially very powerful help in finding end points of what might appear to be 'patterns' in the audio signals (Jusczyk, 2000). For that reason the MFCC representation has been enriched with an estimate of the pitch. Physics-based prosody parameters as well as hand coded accent locations in the speech utterances in the Year 1 databases have been used to investigate the added value of prosody in pattern discovery. So far, these results have been somewhat disappointing, both with the non-segmenting and segmenting approaches to structure discovery. Focusing structure discovery on stretches of speech that contain accented syllables does not result in faster emergence of more powerful internal representations. Almost all structure discovery experiments showed extremely high performance figures with the MFCCs as the only acoustic input. In such a situation it would be extremely difficult for any additional type of input (such as pitch) to have a large effect. It has been suggested that, indeed, the statistics-based bottom-up discovery of acoustic patterns that can be related to references to the environment can succeed without the help of an additional device that focuses attention on the most salient parts of an utterance. The intrinsic salience of the (spectral) properties may be sufficient in its own right. Also, it has been suggested that the role of *prosody* in speech processing comes later, and in different forms. As soon as the learning infant has learned that it may make sense to segments the speech signal in some way or another, prosody may come in helpful in languages where the large majority of the words (meaningful stretches of sound) is characterized by systematic stress patterns (e.g., word stress (almost) always on the first syllable). However, in the experiments performed so far, which were not aimed at the discovery of 'words' in the linguistic sense of the term, systematic stress patterns could make very little –if any- contribution.

Ideally, all experiments should reflect a setting in which a learning agent interacts with a care giver and by virtue of that interaction acquires communication skills, including language proficiency. Fig. 3 gives a schematic representation of the general setting of the learning experiments that have been conducted. The top panel shows the speech corpora from which training utterances are selected (top right hand side). The box labeled 'Experiment design' determines the order in which utterances are taken from the corpora and the number of utterances that are selected as a single set. The selected utterances are made available to the Carer via the 'Stimulus list'. In most experiments conducted so far the Carer offered the utterances in the Stimulus list to the learning agent in the exact same order and manner as determined by 'Experimental design'. In future experiments the Carer will be made more

independent, so that s/he will be able to select utterances on the basis of the response of the learning agent.
The lower part of the schema in Fig. 3 represents the interaction between the Carer and the Learner. Basically, the Carer offers an utterance to the Learner, who then processes this new stimulus. When processing is finished, the Learner will respond. So far, the response options for the Learner are limited to selecting one or more (in the multi-keyword input utterances in the Year 2 database) objects that are supposed to be referenced in the input; alternatively, the response can be NIL. As said above, the implementation of the simulation environment provides for the option that the Carer selects the next stimulus on the basis of the Learner's response, but so far this option has not been used much.

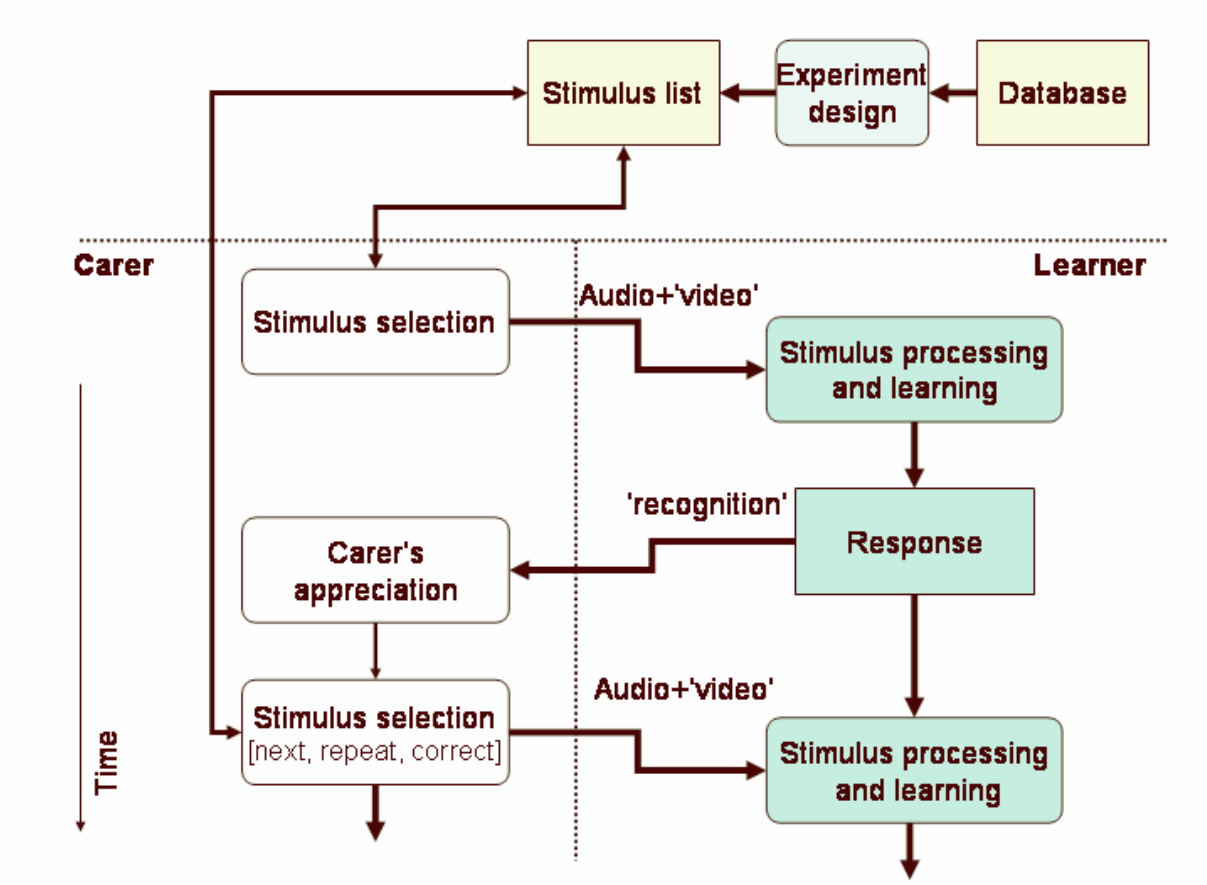

Figure 3 Schematic overview of the learning process. The vertical axis represents time. In a communicative loop, Carer and Learner exchange messages.

The exact meaning of 'Stimulus processing and learning' depends very much on the details of the experiments. For example, in the first stage of the initial experiments with NMF in batch learning mode 'processing and learning' meant just storing the inputs in memory (it is left undecided whether this should best be pictured as sensory store, or rather as working memory). Only after a certain number of utterances were available was the first NMF decomposition applied. Until that moment responses of the Learner would be meaningless. However, after the initialization of the NMF structure matrix subsequent input utterances can be mapped onto one of the objects to be learned.
As already said before, the NMF approach has recently been adapted to enable incremental learning. In this mode the first decomposition is attempted after a small number of utterances has been offered, and subsequent utterances can be used to update the NMF matrices. A similar incremental processing strategy has been developed for the DP-Ngram approach.

It goes without saying that 'processing' as well as the representations in memory are different for the different approaches to discovering structure.

4.1 NMF for discovering 'words' in continuous speech

Non-negative Matrix Factorization (NMF) is a member of a whole family of approaches aiming at the discovery of 'structure' by using a specific decomposition technique. NMF is a general mathematical technique for decomposing a large matrix that only contains non-negative numbers into two smaller matrices, also comprising only non-negative numbers, in such a manner that one of the resulting matrices can be considered as representing 'basic' structural elements and the other as the degree to which these basic structural elements add up to form a given arbitrary observation. During the learning phase NMF builds the representations of the structural elements from scratch. Thus, NMF fulfills one of the basic requirements in ACORNS: it does not impose meta-linguistic knowledge upon the learning process.

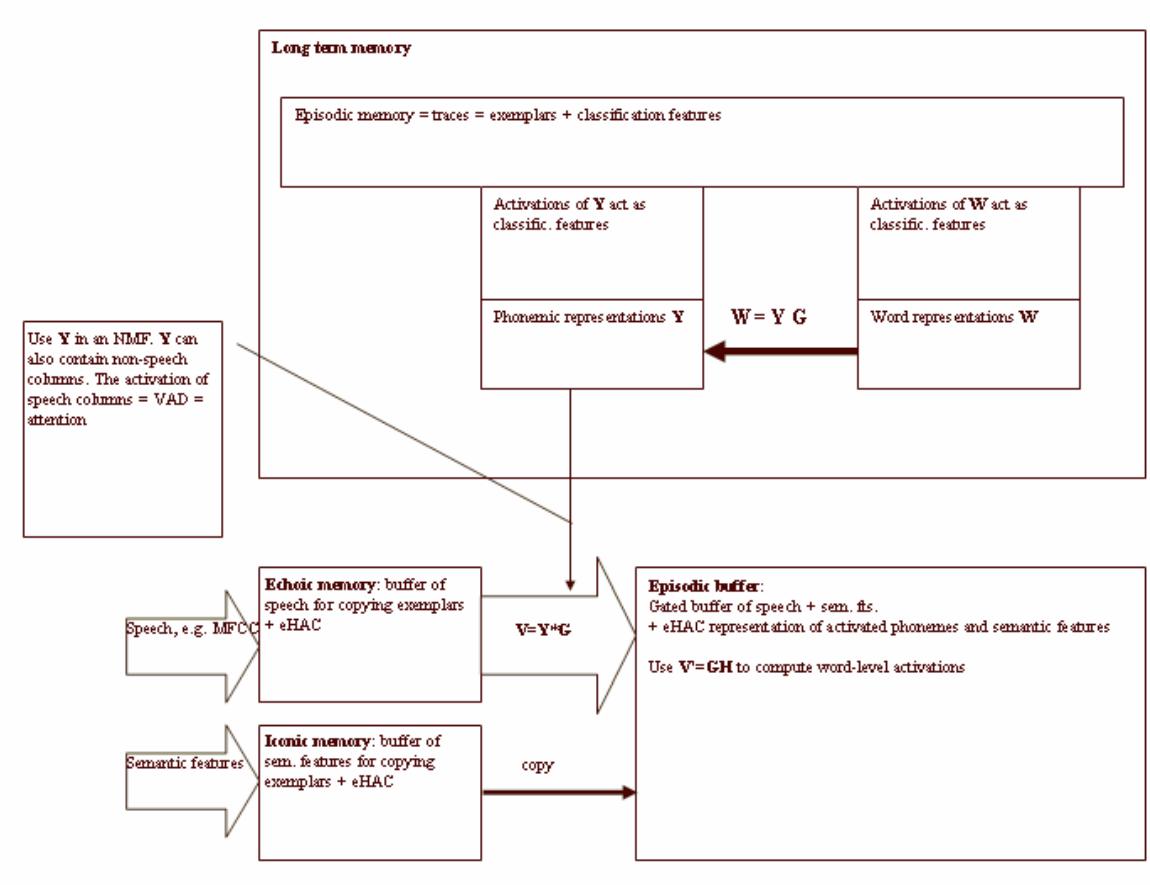

Figure 4 Correspondence between NMF and the general memory architecture.

Still, there may seem to be a big gap between NMF as a mathematical device on the one hand and the memory architecture sketched in Fig. 1 and the learning process in Fig. 3 on the one hand and the operations of NMF on the other. To narrow this gap Fig. 4 explains the way in which NMF can be mapped onto the general framework suggested in Fig. 1. In interpreting Fig. 4 it must be taken into account that the actual implementation of the architecture of Fig. 1 would require many decisions, some of which might run counter to the process depicted in Fig, 4. However, we are convinced that the scheme shown in Fig. 4 is compatible with at least some plausible implementations of the architecture in Fig. 1.

NMF (and almost all other extant algorithms for structure discovery) require that 'objects' (in our case speech utterances) are mapped onto a numerical representation in the form of a vector of fixed length. We have designed such a mapping: the histogram of co-occurrences of acoustic events (HAC). Speech is represented as an ensemble of ordered (in time) acoustic events. In our initial experiments, these were the detection of phones (Stouten et al., 2008; Stouten et al., 2007). Later, they were plainly the observation of typical speech spectra (Van hamme, 2008a). The HAC-mapping is then the accumulation of the number of times any combination of acoustic events occurs in a speech utterance. This utterance-level histogram is clearly affected by the words it is composed of. The NMF will now factorize a collection of such utterance-level histograms into word-level histograms, which form the learned internal representations of words.
Grounding of the learned internal representations is achieved by jointly estimating the co-occurrence of the acoustic events with the occurrence of events in the other modalities (in the Year 1 database, these are crisp keyword tags simulating the visual channel). Hence, we obtain internal word representations that are truly cross-modal and that link information in different modalities to the extent that it is possible to predict the feature values of the visual modality (i.e. keyword tag) from the observed acoustics.
The NMF paradigm allows to learn and subsequently recognize words in a sentence. However, it is a detection type of response, in which an unordered set of words are activated by the recognizer. Since word order is important in language, we extended the method to also estimate the word position within the analysis window of the recognizer, and hence order the words (Van hamme, 2008a).

The NMF-framework has shown to be a powerful approach to word acquisition in which information across modalities can be exploited to build integrated internal representations. Acoustic and semantic information at different time scales can easily be integrated (Van hamme, 2008b) and the learning was made incremental in the sense that learned internal representations can be updated based on a single utterance. Apart from acquisition, it also allows to build a bottom-up speech recognizer, where words are activated from the acoustic evidence without, like in HMMs, the need for maintaining tens of thousands of search hypotheses nor is there a need to sharply segment the utterance into words or subword units (Van hamme, 2008b). Hence, *activation-verification* recognition framework is in place. Thus far, the *verification* component is limited to checking if the activation of words is sufficiently strong and consistent over time, rather than confrontation with learned evidence.
To show reuse of learned representation, a hierarchical model of speech was learned in which the acoustic events used in the HAC-model were also learned with NMF. However, top-down learning of reusable phoneme-like units by analyzing commonalities in learned word models has thus far always lead to significant losses in accuracy.

4.1.1 NMF-based experiments for learning to understand keywords

4.1.1.1 Aim of the experiments

In this section, we specifically discuss experiments designed to investigate the learning curve of an NMF approach depending on the settings of the multiple parameters that can be specified. The learning results investigated include the accuracy of the interpretation as provided by the learner, as function of the number of stimuli presented, the sensitivity of the internal representations in terms of the amount of learning material, and the speaker-dependency of the internal representations. Experimental parameters include the order in which utterances are offered, the number of utterances that are stored before an NMF decomposition is attempted, and the way in which the 'visual' information is encoded. To that end, about 20 different experiments have been carried out during the first 18 months of the project (more details can be found in ten Bosch et al. (2008a) and ten Bosch et al. (2008b).
Interestingly, these experiments do not differentiate between training and test set, as is mostly done in automatic speech recognition. Instead, the entire database of 8000 utterances is processed in an utterance-by-utterance manner. Each utterance-tag pair is presented only once. Learning therefore takes place by 'remembering' the characteristics of the utterances and associations that have been observed.

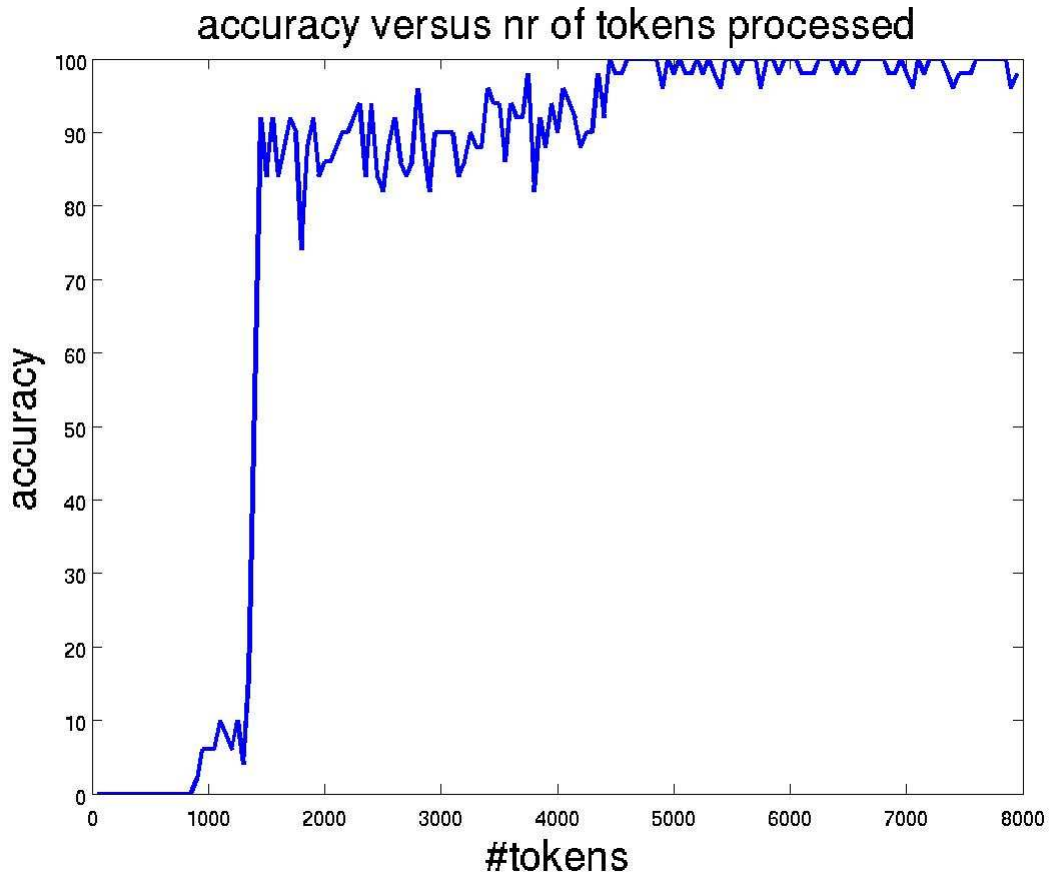

Figure 5 Results of the learning algorithm. The horizontal axis displays the tokens; in total 8000 tokens are presented. The vertical axis shows the accuracy as measured on the most recent 50 utterances.

4.1.1.2 Method

The learning algorithm is based on the assumption that the internal representation is updated as new stimuli are processed. The update rate is one of the parameters to be determined. At a certain point during a training, a new stimulus (utterance + tag) is presented. This stimulus is processed by the perception module, of which the outcome is stored into the sensory store, and from there into the short-term memory. The learner then attempts to interpret the new utterance in terms of the stored representations in the long-term memory. The aim to improve the interpretation of each unseen stimulus in terms of what it knows at that moment is the ultimate ***learning drive*** of the learner.

4.1.1.3 Results in relation to automatic learning - acquisition and ASR

The outcome of all experiments can be summarized as follows.

Firstly, the experiments show that it is possible to build internal representations by processing HAC representations of input utterances by means of NMF decomposition.
Secondly, internal representations are dependent on the speaker. This is shown by the difference between learning curves obtained in a randomized speaker setting and a speaker-blocked setting of the learning experiment. In the speaker-blocked setting, internal representations are built for one speaker, and must be adapted to the next speaker in order to obtain the same performance. In Fig. 5, the performance is shown in the case of randomised ordering of Dutch stimuli. The horizontal axis displays the tokens; in total 8000 tokens are presented to the Learner. The vertical axis shows the accuracy as measured on the most recent 50 utterances (approximating the 'instantaneous accuracy'). In the beginning, no internal representations are built. As a consequence nothing is correctly

recognized and the accuracy is zero. Fig. 6 is similar to Fig. 5, the difference being that the 8000 stimuli are presented in speaker-blocked fashion (female 1, male 1, female 2, male 2). The drop in accuracy for each new speaker shows that representations are speaker dependent if stimuli are presented speaker block-wise, where the blocks contain enough utterances (in this case 2000).

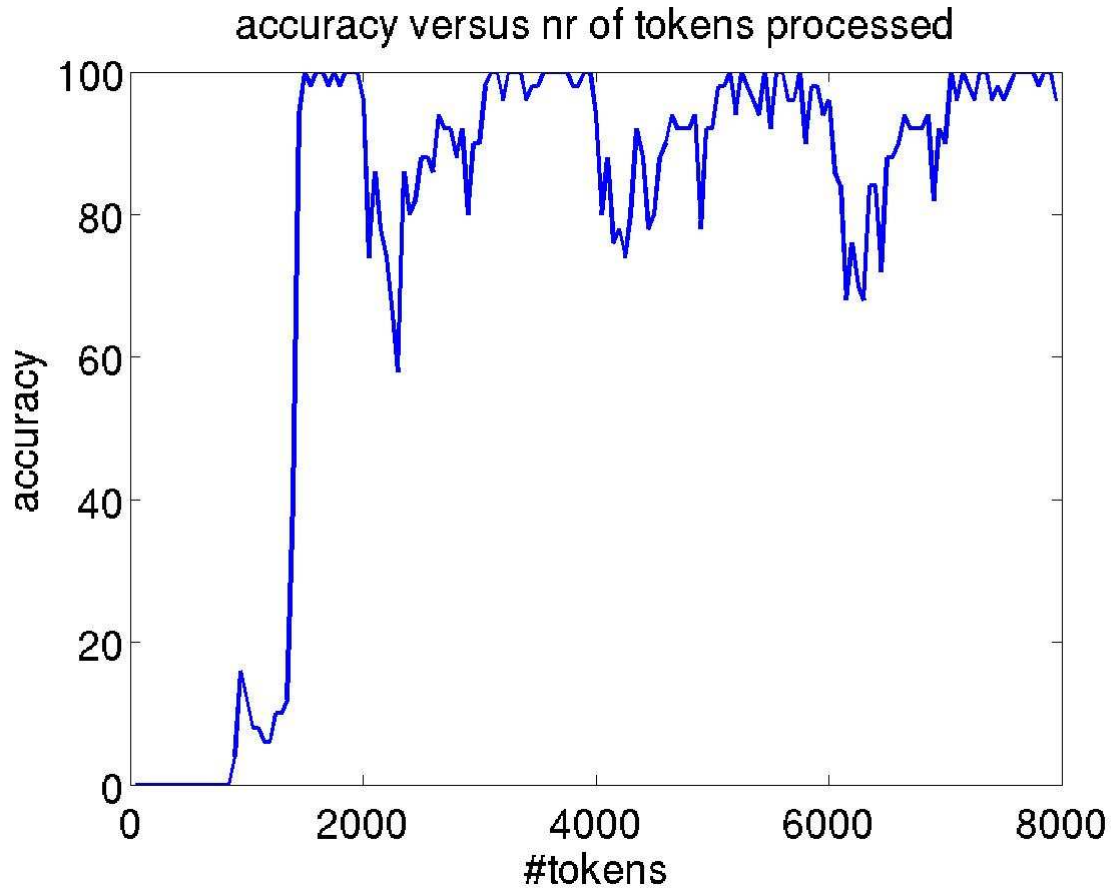

Figure 6 Results of the learning algorithm. Stimuli are presented speaker-block-wise.

Thirdly, learning results are sensitive to a specific set parameters that determine the way how the learning takes place. Interestingly, these parameters have a direct cognitive interpretation. We identified four parameters:

a) the amount of material (stimuli) before internal representations are built.
b) The amount of recently observed stimuli that us used to update existing representations or hypothesize new representations
c) The number of times stimuli must be used internally for this update
d) The 'eagerness' with which updates take place.

Fig. 7 shows the dependency of the learning algorithm in terms of the amount of data used for updating its internal representation. For clarity, the learning curves are shown for the 2000 utterances of the first speaker only. The abbreviations 'nsbt' and 'ml' refer to 'number of stimuli before training' and '(internal) memory length', respectively. Essentially the figure shows the sensitivity of the performance of the learner as a function of the size ('ml') of its working memory. For the update of internal representations, the learner should take into account the information gathered over the last 500 utterances to obtain an eventual performance of beyond 90 percent accuracy.

4.1.1.4 Ideas for the final year

The HAC-model will be made cognitively more plausible by using a memory model that exploits forgetting to order the recognized items in time.

The vocabulary size will be extended by incorporating more acoustic information, i.e. features at different time scales produced by WP1.
We will further explore the hierarchical organization of the learned internal representations.
In the final year we intend to develop theories and experiments on how *abstractions* can emerge. We still do not know exactly how the concept of 'abstraction' should be interpreted. Using the Year 1 and Year 2 databases, we have conducted several experiments related the internal structure of the space of internal representations. One of these experiments was quite revealing in the sense that 'abstraction' on a certain level could be the result of a more efficient use of representational space when the collection of stored representations gets too crowded on a lower level. In this case, abstraction is equivalent to grouping.
This directly relates to the use and implementation of hierarchies in the learning algorithm.

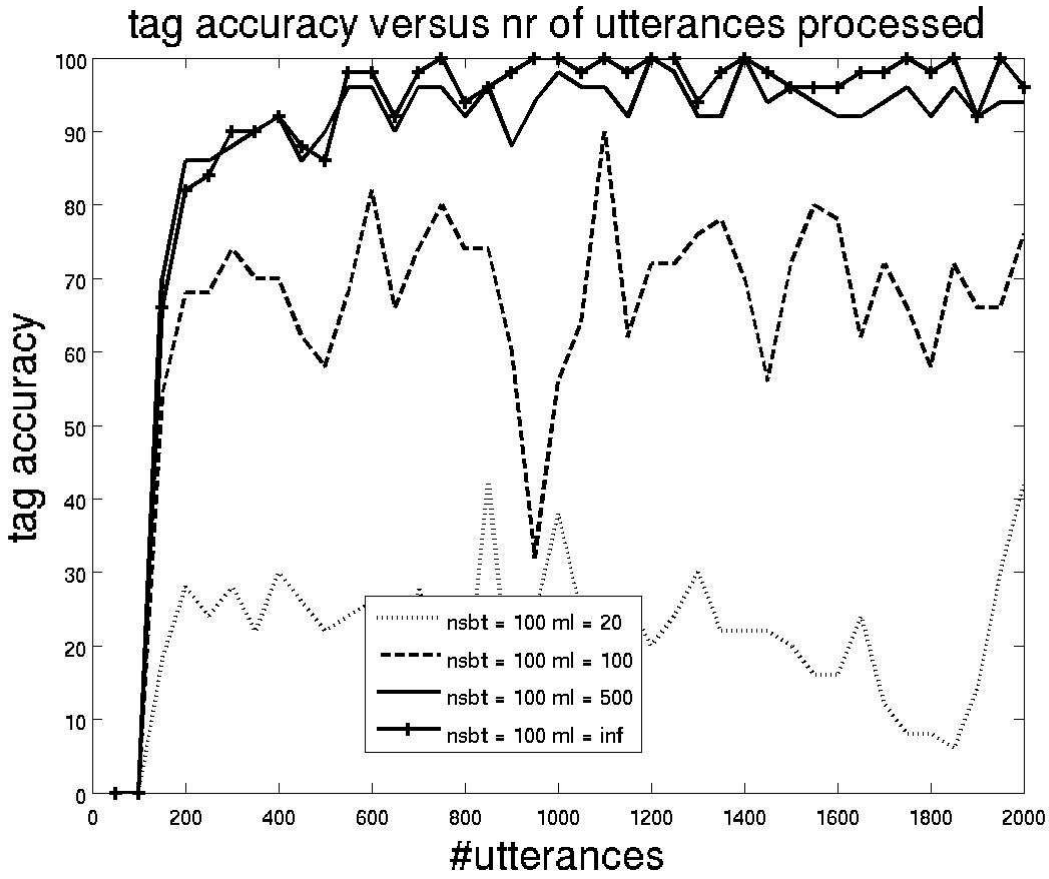

Figure 7 This figure shows the dependency of the learning algorithm in terms of the amount of data used for updating its internal representation. The abbreviations nsbt and ml mean 'number of stimuli before training' and 'memory length', respectively.

4.1.1.5 Questions to SAC

Learning reusable phoneme-like internal representation from an acquired vocabulary has so far remained unsuccessful. For practical reasons, we have limited ourselves to deriving such representations from a vocabulary of about 400 words. Is there any evidence from human word acquisition experiments or from medium-vocabulary ASR that a phonemic organization could emerge from such a small vocabulary? Can we discern if the phonemic/phonetic structure in speech perception is a purely top-down learning process (i.e. observing that the representation of a vocabulary can be simplified by phonemes), a purely bottom-up (i.e. first observing that there are recurring acoustic events, phones and subphonetic events, that are combined to words) or a combination of both ?

One of the issues that remain is the question to what extent a word discovery method is able to explain results of a certain type of psycholinguistic experiments that show how abstraction explains the flexibility of internal representations. In the learning algorithm, representations enter into a competition and the boundaries between representations, once formed, are flexible and updated all the time during training. This adaptive behavior is common in both the explanation of psycholinguistic experiments and in the explanation of the competition between internal representations. So it is plausible to serve as common ground for a more extensive set of experiments.
Another question relates to hierarchy. Is hierarchy the result of a more efficient use of representation space? And if so, which process determines the construction of and the levels in this hierarchy? Another, closely related, issue (that will come back in virtually all other approaches) is to what extent the variation that is pervasive in speech signals can be accounted for on a single layer, if one does not want to impose some kind of structure and units a priori. Accepting units that are defined a priori opens the possibility to learn probabilistic models for those units. In a sense, these putative units can be considered as an extra layer in the architecture. Is it at all reasonable to assume that structure can be discovered in data as variable as speech in a mono-layered architecture?
A third question relates to the data used in the Y1 and Y2 database. In the consortium, we have discussed the pros and cons of this type of 'artificial, semi-spontaneous' real-speaker data at length. The rationale of using this type of data was to keep away from the fully artificial nonsense syllable sequences (Saffran et al, 1996) on the one hand and the 'found speech' as observed in realistic carer-child interaction. The set-up that we settled on in ACORNS is inspired by discussions with Elisabeth Johnson when she was still working in Nijmegen.

4.2 Multigram

The multigram approach is another approach that is being explored in the ACORNS project. The multigram concept is especially useful for detecting recurrent patterns in sequences of symbolic entities.

4.2.1 Aim of the experiment

The aim of this experiment was to see whether the multigram algorithm can be used for word acquisition from spoken utterances that are accompanied by information from other modalities. In the original multigram algorithm (Deligne and Bimbot, 1997), symbolic input is explained by a set of units, multigrams, that emit symbolic strings stochastically. In our (and their) setup, each multigram is modeled by a Hidden Markov Model (HMM). The set of multigrams, their topology and parameters need to be learned. We have extended the multigram learning algorithm to cope with ambiguity, i.e. the input is not a string of symbols any more, but a lattice of symbols that describes a large collection of possible input symbols. Secondly, we have designed a method to link the discovered patterns (multigrams) with the information in other modalities.

4.2.2 Method

We have performed two main sets of experiments.
In the first one, on the well-known TIDIGITS database, we wanted to forego the requirement that patterns are learned without any prior knowledge and rather focus on the pattern discovery itself. In other words, we have assumed that the phone inventory has already been learned before starting to discover word-sized patterns of phones. To this end a conventional HMM-based acoustic model was used to create a phone lattice for each utterance. By making use of our extended algorithm with lattice input, a self-discovered set of multigrams representing word-like units was derived from the train set. The utterances in that set were then segmented into these multigram units and a statistical mapping between the segmented 'words' and the multimodal information was performed.
Finally, by segmenting the utterances in the *test* set and guessing the multimodal information i.e. the 'meaning' of each utterance, a final score could be determined by comparing this guessed sequence of multimodal tags with the actual sequence of multimodal tags present in each utterance.

In a second set of experiments, we wanted to avoid using the prior knowledge of phones. We applied the multigram algorithm directly at the signal level, i.e. to vector-quantized spectra and their velocity and acceleration features, using the ACORNS year 1 database. Each feature stream generates 100 symbols per second, a much higher symbol rate than at the phone level. There were 150 different static labels, 150 different velocity labels and 100 different acceleration labels. Initial HMM's representing word-like patterns were derived from the static stream. The intention was to train these on the train set using all three streams conjointly, much like the way it is done in the training of discrete density HMM's.

4.2.3 Result in relation to automatic learning acquisition and ASR

The results of the first attempted experiment on TIDIGITS were very promising. The results show a significant improvement of taking lattice input into account as well as showing global success discovering and grounding the vocabulary from phone-level input. All eleven words present in the database showed up as separate models in the self-discovered set, together with a small number of distorted versions of some of these models (e.g. “wA” as a distortion of “wAn”) and a couple of garbage models. Mapping these HMM's to multimodal tags and using them to do recognition on the test set yielded tag recognition rates of around 90%. Interesting to note is that the recognition rate when using complete lattices yielded an absolute improvement of approximately 3%, compared to the result when using only the best path through the lattices, giving a tag recognition rate of approximately 93%.

The results in the second experiment, however, were a lot more disappointing. The learning algorithm starts off with an inventory of symbol sequences that occur sufficiently frequently and transforms them into an HMM per multigram. The parameters of the HMMs are re-estimated and the least probably multigrams are pruned. This iterative process is repeated. For VQ-label level input, it proved virtually impossible to determine an acceptable initial set of HMM's. Because of the increased number of symbol identities (150 VQ labels vs. about 40 phone labels) and because of the higher symbol rate, the emission statistics of a multigram become exponentially more complex: much longer patterns with more variation need to be searched for if we want them to model word-like units. Moreover, a lot of variability is present in the label stream that is not observed in a phonetic transcription, such as speed, intonation,... On top of that, asynchrony of the spectral analysis window and the pitch instants can cause alternating sequences. All this causes the stream of VQ-labels to be a sequence that is too erratic to determine stable patterns from. The multigram algorithm relies on a prior probability for each of the patterns in the set, which is initialized by the relative number of occurrences of that pattern in the input. No such probability can be determined if every pattern of interest occurs only once or twice.

Another observation is that the multigram algorithm needs to explain the complete input sequence in terms of multigrams. Hence, it has difficulty to cope with new words or input that is only partly known, which contradicts with human keyword spotting abilities.

4.2.4 Ideas for the final year

Due to its apparent lack of robustness against variations in speech, the pursuit of a full fledged language learning agent based on the multigram algorithm was, at least for the time being, discontinued. Our only hope to overcome the difficulties would be to make the learning process hierarchical, in which shorter (e.g. phone-sized) units with consequently less variation are discovered first. Word-level patterns would then be learned from the output of this layer.

4.3 DP-ngram

DP-ngram is a technique that allows the discovery of recurrent patterns directly by comparing pairs of acoustic utterances. In contrast to the multigram method discussed above, the input is sub-symbolic rather than symbolic. The DP-ngram method hypothesizes templates that are stored in order to be reused later. There are substantial similarities with the ideas underlying ‘episodic’ speech processing.

4.3.1 Aim of the experiment

The experiments carried out using the DP-ngram algorithm show that it can successfully segment speech in an unsupervised and incremental manner. Results show that even with a limited memory of past utterances it is still possible to exploit the statistical regularities in speech to discover word-like units. Similar to the NMF experiments in section 4.2 there is no separate training/test set and utterances are processed sequentially.

The algorithm aims for cognitive plausibility by beginning life with no prior knowledge of any language and carrying out segmentation in an unsupervised and incremental fashion.

4.3.2 Method

The DP-ngram model is able to segment speech, directly from the acoustic signal; automatically segmenting important lexical fragments by discovering 'similar' repeating patterns. Speech is never the same twice and therefore impossible to find exact repetitions of importance (e.g. phones, words or sentences). The use of dynamic programming techniques allows this algorithm to accommodate temporal distortion through dynamic time warping. Traditional template based word spotting algorithms using dynamic programming would compare two sequences, the input speech vectors and the word template, and penalise insertions, deletions and substitutions using negative scores. Instead, this algorithm uses quality scores, positive and negative, in order to reward matches and prevent anything else, resulting in longer more meaningful sub-sequences.

The algorithm is also able to create continuously evolving internal representations of keywords by exploiting the cross-modal statistical regularities in the input stimuli (acoustics + visual tag). From the very first utterance we test the models internal representations of the keywords by asking it to predict the tag of every incoming utterance.

As an incremental process the DP-ngram model only compares the current utterance with a set number of (earliest/latest) past utterances, dramatically increasing processing efficiency at the cost of decreasing the total search space. However, by re-using internal representations we can increase the search space by calling up segments from older utterances stored in the long term memory.

4.3.3 Result in relation to automatic learning acquisition and ASR

Results show that there is enough information in the input to build internal representations of important lexical units even with a very limited storage capacity for processing incoming utterances. It is also apparent that the re-use of internal representations allows the keyword hypotheses to become more accurate and stable at a faster rate by discovering more exemplar representations that would have been outside of the incremental search space.

4.3.4 Ideas for the final year

There are many possibilities for further research and experiments:

a) The addition of prosodic features (rhythm & pitch) as a segmentation and/or attention aid.
b) Hierarchically structuring the input stream with the discovered lexical units would allow us to analyze the relationships between units.
c) Create variance models of the discovered speech units. Will sub-word units like phones emerge as the most efficient units for speech? Variance models of the units could be constructed to see if they exhibit phonetic categorization properties and evidence for native-language neural commitment.

4.3.5 Questions to SAC

The DP-ngram model stores a list of episodic segments for each internal representation of important lexical units (e.g. key word). A major question that has arisen is how to cluster them to find a general/ideal representation. Does an infant use an accumulation/average of all heard exemplars of each unit or a single most ideal episodic representation? Are there computationally efficient methods for combining cluster centroids with strategically chosen exemplars, or alternatively, for keeping some

representation of many exemplars in memory but organized in such a manner that each new input only activates a small subset of those exemplars, so that the computational effort in the search can be kept within feasible bounds?

4.4 Computational Mechanics Modeling

For Computational Mechanics Modeling the algorithm known as Causal State Splitting and Reconstruction (CSSR) has been explored as an approach to discovering structure without imposing a priori units. Similar to the Multigram method, also CSSR takes symbolic sequences as input. That means that this method may be useful AFTER some preprocessing has converted the 'raw' acoustic input into a symbol sequence (or lattice).

4.4.1 Aim of the experiment

The aim of the first year ACORNS experiments was to test the suitability of the so-called Causal State Splitting Reconstruction algorithm, or CSSR (described in the next section), for speech recognition and language acquisition. Since the algorithm appears not to have been tested much in this setting, the first experiment was to apply CSSR to a simple, speech-related dataset in order to study the properties of the learned representations. For this part the experiments used a text transcription of the Swedish first year ACORNS database recordings. Because the algorithm expects input from a small set of discrete symbols, the data was converted to a word-level sequential symbolic representation by considering each unique word in the data as one symbol.

A second goal was to assess the behavior of CSSR on more realistic data, including complications such as noise. To this end, another dataset was generated, similar in character to the symbolic representation of the ACORNS recordings but also featuring low-probability symbol substitution noise.

4.4.2 Method

The CSSR algorithm is a convergent procedure for learning the so-called causal state representation of a stationary stochastic process from empirical data. The causal state representation is a minimal sufficient statistic for predicting the observation sequence; the states contain precisely all information from past observations relevant for predicting the future, and nothing more. The causal state set also has a natural interpretation as an HMM but, as the states are explicitly composed of strings of symbols, the current state can now be uniquely identified from the available sequence of observations.

Unlike traditional ASR with HMM methods, CSSR performs unsupervised pattern discovery, not just recognition of previously learned patterns. However, at present, CSSR requires the data to be sequences of symbols from an alphabet of small size, such as phonemes. The algorithm may not converge if the number of causal states is not finite. There are also two user-set parameters.

4.4.3 Result in relation to automatic learning acquisition and ASR

Results from the symbolic representation of the Swedish ACORNS data were encouraging. Despite the limited amount of data, CSSR learned a near-perfect automaton representation of a stationary stochastic process to generate the observed data. Each state typically represented a specific word or position within one of the carrier sentences.

When applied to the noisy data, on the other hand, the algorithm failed to converge on a limited set of causal states. It appears that the further back the CSSR algorithm looks at the data, more information comes to light that affects future behavior. As the algorithm is intended to capture all information relevant for prediction, these differences count even if their influence is small enough that a human would label them as noise. The algorithm thus discovers a very large number of distinct possible predictions for the future, each of which has to be represented by its own state.

4.4.4 Ideas for the final year

In response to the high sensitivity to noise, an extension of CSSR is in development. The extension aims to reduce the complexity of the learned representations and increase robustness by adding a user-set resolution level to the state splitting decisions within the algorithm. Only features indicative of large differences in future behavior are to be learned, and therefore be represented as distinct states, while less influential differences should be left to one side.

A possible application of CSSR within the ACORNS framework would be to learn prosodic and phonic input streams from ACORNS sentences in parallel, and later reconstruct such prosodic information from phones alone, for segmentation purposes.

Besides CSSR with resolution, other Markov model learners where states are represented as sets of observed strings are being considered, both as alternatives and for comparison.

4.4.5 Questions to SAC

Unlike some approaches within ACORNS, the HMM-style automaton output of CSSR is quite reminiscent of classical ASR. Where in the language acquisition and speech recognition process would discovery and use of such HMM pattern representations be biologically plausible?

4.5 Automatic Segmentation

4.5.1 Aim of the experiments

Phonological parsing of continuous speech may be based on event detection or phone-like segments (Carlson-Berndsen, 1998). Since phonemes are the smallest units that are able to affect distinctions between words a bottom-up based ASR system should utilize this kind of representation at least on some level. The phone-like segments also provide a good way to represent relevant information in a compact manner. We have not confined our studies only to methods dealing with segmentation but have also continued work towards classification (clustering) of the segmental information based on our own version of *incrementally learning vector quantization (ILVQ)*. Our general goal has now been set higher than *signal patterning* or *pattern discovery* as mentioned in the ACORNS Technical Annex. We are developing a novel bottom-up architecture (agent) able to learn spoken messages from examples that have rich internal representations and methods to cope with the real external world. This method can be tested as one proposal among others coming out of ACORNS.

Our experiments so far can be catalogued as follows:

a) Performance tests and studies of the segmentation algorithm
b) The effect of spectral representations on performance in noisy situations (FFT vs. MFCC)
c) Tests with incremental clustering (ILVQ) and preliminary ASR tests based on this model
d) Tests on the effects and meaning of a simple attention mechanism
e) Comparisons of ADS vs. IDS speech in Finnish and in Swedish

4.5.2 Method

In our framework speech is first segmented into phone-like units using the blind segmentation algorithm developed during Y1. The algorithm detects internally (spectrally) coherent regions of speech and places segmental boundaries to create units. Segmental categories are then obtained incrementally using these segmental units as input to the classification process. After exposure to a sufficient amount of speech, the statistics of the categories consolidate and the speech signal can be described as a sequence of activations of these categories in a systematical manner. The output of this process for each incoming speech signal is a sequence of categorical labels. Statistical methods can then be applied to the sequences in order to discover structures, e.g., sub-word or word-like units.

We have performed a simple word learning experiment where transitional probabilities of adjacent segments (~phones) were tracked in the presence of multimodal input simulated by a visual tag. The

system was able to differentiate between the keywords found in the Y1 corpus (Räsänen, Laine & Altosaar, 2008) and to locate these keywords temporally in the utterance with good accuracy.

Experiments with an external attentional mechanism were performed, in which occurrences of segmental units during the spoken keyword were emphasized over other phonetic content in the input. The idea was to test whether the prosodic aspects of input, or some other salient events in the environment that pinpoint the keyword in some context, could help the system to learn keywords. It was determined that at this still relatively low level processing level external attentional mechanisms are not necessary since all of the acoustic information can be processed in a bottom-up manner and familiar patterns will be discovered in the input despite a biasing focus to a specific part of the stream.

Spectral and temporal properties of IDS and ADS speech were also analyzed and reported in (Räsänen, Altosaar & Laine, 2008). Using the Year 1 corpus as input, we tested whether our bottom-up learning agent was able to differentiate between these two speech types by their spectral content. The system learned separate segmental category models for ADS and IDS speech and tested which model reacted better to speech segments in new input. With all speakers, the ADS/IDS speech type recognition rate was above chance (62 - 95 % correct).

4.5.3 Result in relation to automatic learning acquisition and ASR

The results show that it is possible to bootstrap self-supervised word learning using segmental level (coarse) statistics of the input in a purely bottom-up manner. This replicates several findings from behavioral literature (see, e.g., Saffran et al. (1996), Werker & Tees (1984) and Smith & Yu (2008). An unsupervised bottom-up path from acoustics to complex internal linguistic representations by statistical analysis has been illuminated from several aspects and it will be interesting to see what kind of representations can be formed in this manner.

We have also developed a novel method for unsupervised speech segmentation and also clarified and unified the evaluation methods for automatic speech segmentation via a new R-value measure (Räsänen et al., submitted). Work has also been performed on novel methods for learning vector quantization that are especially designed for describing phone-like segments in a speech stream. In order to better understand discrete time-series representations of speech signals in ASR and learning, we have analyzed both segmental and fixed-frame approaches for structural description and pattern discovery (also in collaboration with the NMF group).

4.5.4 Ideas for the final year

We will continue to develop the incrementally learning, bottom-up system and to test its performance on all levels. The main quality measure in this development work will be word recognition rates on the ACORNS corpora. Therefore, we require a complete (bottom-up) ASR-system to run these experiments. We will also compare the results with the present, well performing NMF-system. We also hope to develop a deeper insight into bottom-up statistical processing of speech signals and find important links to research findings from infant language acquisition.

4.5.5 Questions to SAC

There is plethora of discussion around the concepts of *bottom-up* and *top-down* processing. In language acquisition and comprehension, at what level of processing does some kind of *top-down* feedback occur and at what point in infant development does this feedback begin? What is the type of knowledge that can be learned from speech that affects the further parsing of speech signals? At what level of processing does this occur (acoustics, sub-word units, words?).

4.6 Maps

4.6.1 Aim of the experiment

This section of the paper describes the development of an attention-gated recurrent working memory model that was developed towards the overall aim of the ACORNS project of producing language and communication skills based on sensory inputs in an emergent manner. Hence this computational model is a small step towards the aim of the project to produce an agent that can communicate. In terms of the stances identified in the examination of computational models of language learnability by Kaplan, Oudeyer & Bergen (2008) this research fits within the statistical stance and the embodied and social cognitive stance. As the model relies on extracting linguistic statistical patterns from speech signals that offer a rich sensory input and is constrained by characteristics of the cerebral cortex. The model combines an approached based on reinforcement learning to differentiate between speech and non-speech signals that can act as an attention mechanism so only speech is introduced into the main recurrent self-organising network working memory model that learns speech through a memory representation.

4.6.2 Method

As can be seen from Fig. 8 in the neural architecture the auditory signal is split into fixed time frames using moving overlapping windows with mel-frequency spectrum values extracted for each frame to represent the auditory signal. Auditory frames are used to train the attention-gating element of the model to perform reinforcement learning to determine if the auditory signal is speech or non-speech. Once the decision is made as to whether an auditory section is speech, this can be used to control the input to the recurrent self-organising map model for learned emergent representation of speech. It was decided currently to use fixed time frame-by-frame approach for the inputs into the attention-gated working memory model as it offers a real-time focus. With this approach it is possible to introduce speech frames into the recurrent working memory model as they are identified by the gating mechanism as they are heard over time. In Fig. 8 the recurrent self-organising model not only receives as input a speech frames, but also the activations of the previous time step of the self-organising network which provides a memory representation of the previous speech frames making up the speech element to produce an emergent speech representation.

4.6.3 Result in relation to automatic learning acquisition and ASR

When presented with speech (from the English ACORNS database) and non-speech (crowd noise) samples the trained reinforcement attention-gating system is able to detect correctly 93% of the non-speech auditory and 80% of the speech frames. The incorrect detection of speech frames by the attention-gating network is due in part to periods in the speech samples where there are no speech sounds for instance between words and so is unlikely to have little impact on the representation created by the recurrent working memory model. The recurrent working memory system creates distributed temporal representations on the upper level self-organising map in Fig. 8 that are associated with specific speech sounds. For instance, the top left hand area of the self-organising map represents the sound ‘S’ at the end or start of words such as ‘matches’, ‘taps’, ‘news’, ‘seen’, and ‘comes’. The top right of map is associated with sounds such as ‘SH’, ‘CH’, ‘JH’ and ‘K’. In terms of the working memory model outlined by Baddeley (1992) the recurrent self-organisation model recreates some of the functionality of the phonological loop as it is able to store and assist in the acquisition of word. As the working memory model is currently concentrating on developing representation of the speech waveform signal the representation created is associated with the speech sounds found within this signals. However this is only a single component of representing speech by the cortex and as such the model will be extended in the future to incorporate visual semantic feature into the emergent representation of speech.

4.6.4 Ideas for the final year

As part of the future work it is the aim to extend the current model by incorporating semantic features. As stated by Pulvermüller (2003) semantic feature also play a role in the representation in a word. For content words the semantic factors that influence the cell assemblies come from various modalities and include the complexity of activity performed, facial expression or sound, the tool used etc. Hence, the aim in the future is to combine the representation of the speech signal with a semantic representation of the word to give a richer representation of the word. The recurrent self-organising representation of the speech signal is to be associated with a representation of the semantic features of the word using a further recurrent self-organising approach at the highest level.

4.6.5 Questions to SAC

When associating a speech signal and visual input what is the role of synchronization between them?

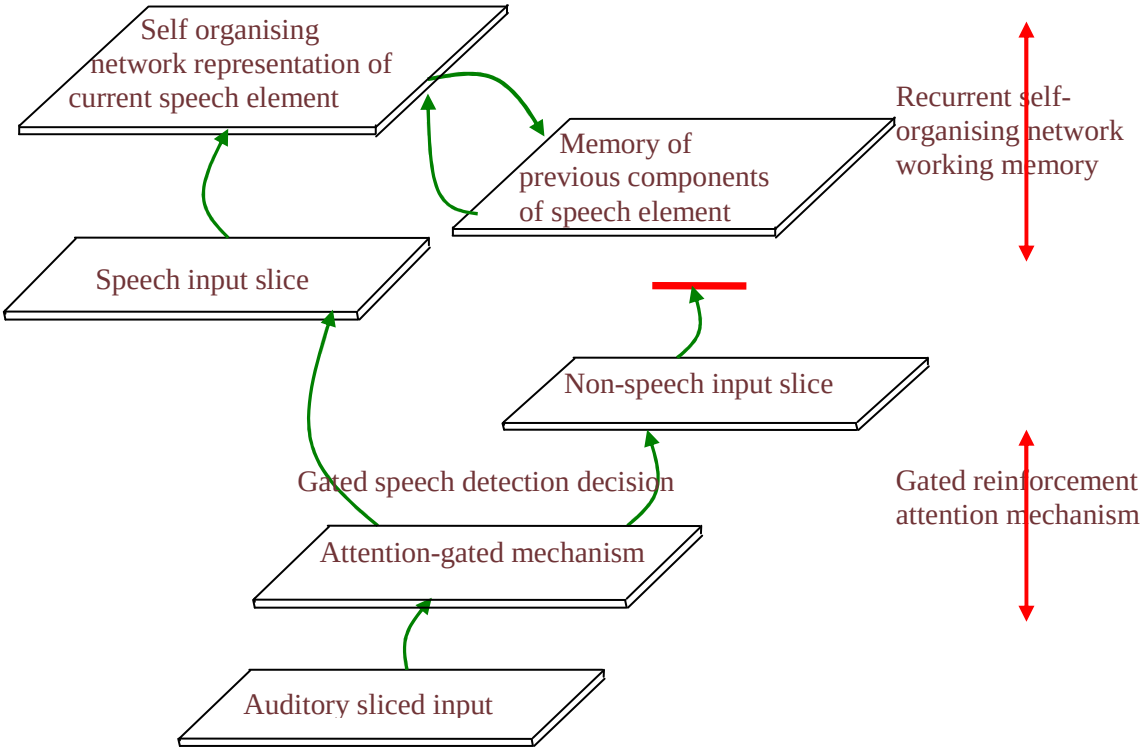

Figure 8 The gated attention recurrent working memory model for emergent speech representation.

4.7 Features that Exploit Auditory Knowledge

4.7.1 Aim

The auditory periphery provides a pre-processing for human speech recognition. To reach human recognition performance, signal features that are not audible are, at least in principle, not needed. More importantly, the human auditory system provides an important indicator of relevant distances between speech sounds. We conjecture that an acoustic feature set with Euclidean distances between sounds that approximate the corresponding distortions indicated by the human auditory system

simplifies the design of a recognition system. Only when a recognizer has an input space with perceptually meaningful distances is the learning of meaningful signal patterns possible. This is particularly relevant in ACORNS since it aims to model an infant who has no prior knowledge of speech sounds.

In this part of ACORNS, our aim is to develop a systematic manner to find a feature set where the Euclidean distance between sounds approximates the distortion at the output of sophisticated auditory models. This work is motivated by the fact that models of the auditory periphery have been improved significantly in recent decades. Until now, these models have not seen significant use in the context of speech recognition.

4.7.2 Method

Consider a vector x describing a speech signal segment. The objective is to find a feature set for which any perturbation ε of the vector x leads to a Euclidean distance that approximates the distortion indicated by the auditory model. Naturally, this has to hold true for all speech signal segments. To measure the similarity of the auditory model distortion and the feature vector distance we can correlate the auditory model distortions and the feature distances found for the ensemble of all speech segments x and all perturbations ε. A higher normalized correlation corresponds to a better feature set.

The use of small perturbations allows us to simplify both the distortion measure of the perceptual model and the Euclidean distance for the feature set. We define quadratic approximations for the perceptual distortion and the feature distance, reducing the computational effort to a reasonable amount.

As a first experiment, we selected a perceptually motivated subset of mel-frequency cepstrum coefficients (MFCCs) from a set of twelve MFCC for clean speech, using the well-known AURORA database. This database consists of balanced sets of clean and noisy utterances. We used a relatively simple auditory model. We checked recognition performance of subsets selected with our method using the HTK recognition system. The full set of twelve features resulted in a recognition rate of 97.4%. Using our method to select the subset of MFCCs, we retained a recognition rate of 96.6% for four MFFCC. In contrast, the recognition rate averaged over randomly selected subsets of MFCCs was 84.2%. Similar results were found for other subset sizes. This indicates that perceptual measures can indeed be used to select a good subset of features.

We are currently working on a second experiment where we optimize the parameters of known features, with the aim to obtain more effective features.

4.7.3 Result in relation to automatic learning acquisition and ASR

In any bottom-up recognition system such as ACORNS it is not logical to select or optimize the features for recognition performance. Features must be selected based on a different criterion. The outcome of this work is a systematic method to find features with Euclidean distances between sounds that correspond to the “perceived distance” between sounds. Thus, instead of using a particular recognition system to find a good set of features, we use the human auditory periphery to select a good set of features.

4.7.4 Ideas for the final year

It is natural to extend the work in two directions. First, various auditory models can be used and their performance can be compared. We are particularly interested in the use of auditory models that show the effect of time-domain masking. Second, the method naturally leads to the definition of new features. We can either use the method to optimize parameters in existing features or to build features from scratch.

4.7.5 Questions to SAC

Are there indications that features that are not perceived by humans are useful in speech recognition?

4.8 Experiments with Semantic Features

4.8.1 Aim

Although the main switch from the usage of rigid tags to the use of semantic features will come with the use of the Year 2 database, first explorations with semantic features were already made during the first year. The primary issue we needed to solve was what sort of cognitive plausible module could provide a good interface between the presented features and the parts of the ACORNS implementation which is concerned with the acquisition of the auditory representations. The function of such an interface would be to take in the perceived semantic features of an object and transform it into activation or probability values of possible concepts. In line with the memory prediction model, we strived to allow this part of the ACORNS implementation to work with minimal a priori knowledge, i.e. the processing architecture should not have any prior knowledge about what concepts it was expected to learn. Further, the architecture should be able to create a representational hierarchy similar to the one observed in the sensory systems of the human brain. While a mapping from low level sensory features to high level sensory features was taken as given (ACORNS is not about sensory processing other than in the auditory domain), we tried to capture the hierarchy from high level semantic features to conceptual units. As argued by Levelt, Roelofs, and Meyer (1999), concepts cannot be represented as feature clusters, but need to be represented as non-decompositional conceptual units. One major reason for this is the hypernym/hyponym problem: if concepts are represented as feature clusters, the conceptual representation of any word has at least all the features of its hypernym. It would therefore be impossible to activate, e.g., the word “daddy” without triggering the word “man” at the same time. However, if concepts are indeed represented as non-decompositional units, at least two other problems remain to be solved (given that the input to the semantic system are perceptual features): (i) explaining the emergence of such units during the acquisition of conceptual knowledge and (ii) preventing the analogue to the hypernym/hyponym problem to occur at the conceptual level, i.e., the problem that the input features leading to the activation of conceptual unit [[daddy]] would also trigger the activation of [[man]]. While the hyperonym problem is not very prominent in the first year database with its limited words, it will already be in the Year 2 database when words like *man* and *daddy*, *bear* and *toy*, or *food* and *apple* are in the vocabulary of almost every child, indicating that the human cognitive system is able to deal with this problem.

4.8.2 Method

In keeping with the general spirit of the memory-prediction model, we tested several biologically inspired learning algorithms, among them: direct Hebbian associative learning, Self-Organizing Maps, biased competitive layers, as well as Restricted Boltzmann Machines. All but the Hebbian algorithm allow the recruitment of units on a higher level to represent higher level concepts.

4.8.3 Results

After testing several architectures, the most appropriate for this task was the competitive-layer architecture. We demonstrated that the conceptual analogue of the hypernym/hyponym problem did not occur in this architecture. The architecture was further able to acquire the correct conceptual units as higher level representations from the presented semantic features without any a priory knowledge in a purely unsupervised manner. Finally, we could replicate overgeneralizations (e.g., the use of the [[dog]] concept for everything with four legs) reported in the literature (e.g., Clark, 1973), one of the most basic behavioural findings related to semantic features during first word acquisition.

4.8.4 Ideas for the final year

One exploratory study would be to use simple images of objects and use the Restricted Boltzmann Algorithm to extract the features from the image. This would reduce the arbitrariness of the selected features and make the visual/semantic process more similar to the auditory process.

4.8.5 Questions to SAC

How realistic is our training situation in comparison to the learning situation of real children. In particular, how much information do children refer about which word in an utterance refer to which object in a scene.

5. General questions to the SAC

From the literature on child language acquisition it is increasingly clear that more than one 'learning mechanism' is involved. There is convincing evidence that infants use statistical patterns in speech (Saffran et al., 1996) In one experiment 8m old infants listened to a 2 minute continuous nonsense sequence of syllables simulating sequences of multisyllabic words ('golabupabikututigolabubikutu'). Infants appeared to be able to distinguish syllable sequences with higher frequencies from 'words' made of lower-frequency sequences, showing that infants are indeed able to detect and use the statistical properties of the speech stream. Later experiments showed that infants are sensitive to transition probabilities (Saffran et al., 1999, Swingley, 2005). The structure discovery approaches in ACORNS seem to be able to find similar statistical structure in real speech signals.

Infants also seem to be able to infer some kind of generalisations from statistical patterns. For example, experiments suggest that infants may detect violations of tokens that do not meet phonological patterns, but the results are not uncontroversial (Marcus Marcus, Vijayan, Rao, and Vishton, 1999, see also Seidenberg, MacDonald, & Saffran, 2003). However, it is not evident what simulation experiments should be performed to arbitrate between alternative interpretations of the behavioural data.

Perhaps the most important thing that we learned in the first two years of the project is that the gap between existing models of memory, representation and learning in the literature on language acquisition on the one hand and computational models that can take real speech utterances as input for learning on the other hand is much larger than anticipated. Due to this gap it is possible to map quite different learning approaches, resulting in quite different internal representations, onto the general models shown in Figs. 1 and 2. One consequence of this is that it is difficult to interpret the different strengths and weaknesses of the approaches that we have explored so far. It is also very difficult to design simulation experiments that directly tap into the questions that still linger in the linguistically oriented language acquisition community.

This raises the following questions

1. What would be experiments that we can do with the techniques and databases that we have developed so far that could answer burning questions of the SAC members
2. What would be the best way to proceed to 250 words, the final goal according to the Technical Annex.
3. Does it make sense to go to 250 words, or are there other issues that can be addressed without a new database (and preferably also without new computational modeling approaches) that are more interesting?
4. Any attempt to go to 250+ words would require some kind of hierarchical (layered) representation, with appropriate corresponding processing. So far we have not managed to define one single processing strategy that is adequate for all types of representations and on all levels (except perhaps in purely connectionist models). It has been suggested that the idea that the seeming homogeneity of the neural fabric should result in the same way of processing at all times and all levels is misleading. Language acquisition literature suggests that new processes come into play as the number (and perhaps complexity) of the neural representations increases.
5. How to exactly address hierarchy and generalisation? In several experiments, we have observed glimpses of some form of generalisation.

- speaker-dependent word representations can be clustered to obtain speaker-independent word representations – this COULD be interpreted as a result of a too crowded set of representations on a lower hierarchical level
- the learner is able to discover the speech style (IDS/ADS) based on many word tokens with an accuracy of about 80 percent – too many to store them individually, so a form of 'abstraction' must have taken place

All **structure discovery approaches** that we have investigated so far appear to be hampered by the large amount of variation in speech signals. To the best of our knowledge there are no structure discovery methods that can operate on semi-continuous inputs, such as MFCC vectors, or other frame-based representations of the signal. The only exception here might be artificial neural networks, but it is questionable whether networks can be trained for such a large task as language acquisition.
It has been suggested that all structure discovery methods that have been developed for discrete symbolic data will encounter serious problems if the number of labels grows too large, and 'too large' might well be anything with more than 50 – 100 elements in the basic set (Lin et al., 2007). If this holds true for all structure discovery methods, what are the consequences for language acquisition? Can the problem be solved, at least in part, by not requiring exact matches between input signals and the representations of a finite set of units? Would it still be possible to mix units of different size (sound, syllable, word, etc.) and combine these for building larger structures on higher levels of the hierarchy?

References

Baddeley, A. D. (1992) Working memory. *Science*, 255(5044), pp. 556-559.

Carlson-Berndsen, J. (1998) *Time map phonology*, Finite state models and event logics in speech recognition, Kluwer Academic Publishers.

Clark, E. V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In T. Moore (Ed.) *Cognitive development and the acquisition of language.* New York: Academic Press, pp. 65–110.

S. Deligne and F. Bimbot, "Inference of variable-length linguistic and acoustic units by multigrams," *Speech Communication*, vol. 23, pp. 223–241, 1997.

George, D., and Hawkins, J. (2005) A Hierarchical Bayesian Model of Invariant Pattern Recognition in the Visual Cortex. *Proc. of the International Joint Conference on Neural Networks 2005*. Montreal

Hawkins, J. (2004) *On Intelligence*. New York: Times Books

Hawkins, J. (2005) Response to reviews by Feldman, Perlis, Taylor, *Artificial Intelligence*, Vol. 169, pp. 196–200.

Jusczyk P. W. (2000) *The Discovery Of Spoken Language* Cambridge, Mass.: MIT Press Ltd

Kaplan, Oudeyer and Bergen (2008) Computational models in the debate over language learnability, *Infant and Child Development*, Vol. 17, pp. 55–80.

Levelt, W.J.M., Roelofs, A.P.A., & Meyer, A.S. (1999). A theory of lexical access in speech production [target paper]. *Behavioral and Brain Sciences*, 22 (1), 1-37.

Lin, J., Keogh, E., Wei, L. & Lonardi, S. (2007) Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, Vol. 15, pp 107–144.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by 7-month old infants. *Science*, *283*(5398), 77–80.

Mountcastle, V. (1978) An organizing principle for cerebral function: The unit model and the distributed system. In: G.M. Edelman and V.B. Mountcastle, Editors, *The Mindful Brain*, MIT Press, Cambridge, MA.

Pfeifer, R. & Scheier, C. (1999) *Understanding Intelligence*. Cambridge: MIT Press.

Pulvermüller, F. (2003) *The neuroscience of language: On brain circuits of words and language*. Cambridge Press.

Räsänen O., Laine U.K. & Altosaar T. (submitted) ---- .*Speech Communication*

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

Räsänen, O., Laine, U.K. & Altosaar, T. (2008) Computational language acquisition by statistical bottom-up processing. *Proc. of Interspeech '08*.

Saffran J., Aslin R. & Newport E.(1996) Statistical Learning by 8-Month-Old-Infants. *Science*, 274, 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52.

Saffran, J.R., Werker, J.F. & Werner, L.A. (2006) The Infant's Auditory World: Hearing, Speech and the Beginnings of Language. In: Damon, W., Lerner, R. M., Kuhn, D. & Siegler, R. S. (Eds.) *Handbook of Child Psychology, Volume 2: Cognition, Perception, and Language*, New York: Wiley, pp. 55-108.

Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2003). Are there limits to statistical learning? *Science, 300*, 53–54

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

Stouten, V. Demuynck, K. and Van hamme, H**.** (2007) Automatically Learning the Units of Speech by Non-negative Matrix Factorisation. In *Proc. European Conference on Speech Communication and Technology*, pages 1937-1940, Antwerp, Belgium, August 2007.

Stouten, V., Demuynck, K. and Van hamme, H. (2008) Discovering Phone Patterns in Spoken Utterances by Non-negative Matrix Factorisation. *IEEE Signal Processing Letters*, volume 15, pages 131-134, 2008.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*, 86–132.

ten Bosch, L., Van hamme, H. and Boves, L. (2008a) A Computational Model of Language Acquisition: Focus on Word Discovery. In *Proc. International Conference on Spoken Language Processing*, pages 2570-2573, Brisbane, Australia, September 2008.

ten Bosch, L., Van hamme, H. and Boves, L. (2008b) Unsupervised detection of words - questioning the relevance of segmentation. In *Proc. ITRW on Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark, june 2008. 4 pages, ISBN

van Doremalen, J. and Boves, L. (2008) Spoken Digit Recognition using a Hierarchical Temporal Memory, *Proc. Interspeech* 2008, pp. 2566-2569

Van hamme, H. (2008a) HAC-models: a Novel Approach to Continuous Speech Recognition**.** In *Proc. International Conference on Spoken Language Processing*, pages 2554-2557, Brisbane, Australia, September 2008.

Van hamme, H. (2008b) Integration of Asynchronous Knowledge Sources in a Novel Speech Recognition Framework. In *Proc. ITRW on Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark, June 2008. 4 pages, ISBN 978-87-92328-00-7.

Werker J. & Tees R. (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavioral Development*., 7, 49-63.