ACORNS – towards computational modeling of communication and recognition skills

Lou Boves[1], Louis ten Bosch[1], Roger Moore[2]
[1]Dept. Language and Speech, Radboud University Nijmegen, [2]Dept. of Computer Science, University of Sheffield
l.boves@let.ru.nl; l.tenbosch@let.ru.nl; r.k.moore@dcs.shef.ac.uk

***Abstract*-In this paper the FP6 Future and Emerging Technologies project ACORNS is introduced. This project aims at simulating embodied language learning, inspired by the Memory-Prediction theory of intelligence. ACORNS intends to build a full computational implementation of sensory information processing. ACORNS considers linguistic units as emergent patterns. Thus, the research will not only address the issues conventionally investigated in statistical pattern recognition, but also the representations that are formed in memory. The paper discusses details of the memory and processing architecture that will be implemented in ACORNS, and explains how this architecture merges the basic concepts of the Memory-Prediction theory with results form previous research in the field of memory.**

I. INTRODUCTION

The conventional automatic pattern recognizers that we have today are all to a large extent based on the same underlying principle: during the training phase they form statistical models to represent the training data in terms of a fixed and limited number of *pre-defined* patterns, and during recognition they search for the sequence of trained pattern that best matches the incoming signal.

This approach has been very successful for the classification of 'physical' objects such as cars, airplanes, cells, chromosomes, or the sounds produced by machinery, even if these objects are observed in conditions that differ mildly from those represented in the training data. In applications such as license plate recognition these statistical pattern recognizers can outperform humans.

However, it is well known that there are domains where conventional automatic pattern recognizers fall dramatically below the performance of biological systems. Automatic speech recognition (ASR) is one such domain (Lee, 2004; Moore, 2005). Advances in hardware, algorithms and data structures have enabled the implementation of large vocabulary, continuous speech recognition (LVCSR) systems and the use of certain voice-enabled dialogue systems. However, existing speech recognition applications are restrictive, requiring that their users follow a strict protocol. The technology is quite fragile, and careful designs have to be adhered to rigorously if the technology deficiencies are to be overcome (den Os et al., 2005). In adverse acoustic conditions, when the mismatch between training and test data is large, the accuracy of today's ASR system accuracy declines dramatically, to the extent that they become unusable, even for cooperative users. When compared with *human speech recognition,* or HSR, the error rates of state-of-the-art ASR systems are an order of magnitude larger, even for rather simple tasks in noise-free environments (Lippmann, 1997; Sroka & Braida, 2005; Wesker et al., 2005).

Apart from the fact that the performance gap hampers the commercial introduction of flexible voice-driven applications, there is also a more scientific issue to be addressed. Due to the nature of statistical modeling and optimization based on statistical principles, today's statistical pattern recognizers may be fundamentally incapable of closing this performance gap (Moore, 2003), and it may well be that new computational methods must be applied that take into account key aspects of human speech processing. Recently, there has been a growing interest in approaches that explore new directions in cognitive modeling, especially in learning and language acquisition. Roy and Pentland (2002) focused on machine learning of words; Werker and Curtis (2005) presented a comprehensive model of human language acquisition, while Maloof and Michalski (2004) focus on incremental learning. The developments in the field of learning in general and language acquisition in particular can be linked with seemingly independent developments, such as the introduction of Hierarchical Temporal Memories based on recent findings in neuroscience (Hawkins, 2004) and the more general trend towards embodied models of cognition (Pfeifer and Scheier, 1999).

The three-year European Union 6th Framework Future and Emerging Technologies project ACORNS (started on December 1 2006) aims at integrating the developments alluded to above, so as to develop a novel approach to speech recognition. We will use state-of-the-art learning models in combination with advanced memory architectures. The project is based on the conviction that the inability of statistical pattern recognizers to approximate human performance is due to the fact that these systems have been designed to emulate theories about the structure of speech and language, rather than to model the purposeful processing of speech for the purpose of communication and, as a last resort, to survive in the natural habitat. This corresponds to the fact that conventional pattern recognizers are trained to discriminate pre-defined patterns that are invariably based on a human-crafted meta-description of the phenomena to be dealt with. For Automatic Speech Recognition (ASR) this means that systems are trained to recognize 'words' that are represented in the form of a

sequence of discrete sounds. However, although such a representation of words may be very convenient for the purpose of linguistic description, it does not reflect the fact that speech production is fundamentally a continuous process (Ostendorf, 1999). In terms of the theory of embodied cognitive science this effectively means that conventional (ASR) systems make an error of frame-of-reference: an observer-based representation of some aspect of the task is confused for the procedure needed to solve the task (Pfeifer and Scheier, 1999).

For humans (and probably al other living organisms as well) 'patterns' (or perhaps more appropriately: "ecologically relevant deviations from randomness") in the sensory inputs are emergent properties (Johnson, 2002) that are learned because of the innate need to associate (inherently variable) sensory inputs to meaningful objects and behavior in the environment. Because the sensory signals corresponding to ecologically relevant entities (e.g., words, phrases, multiword expressions, etc.) are so variable, while 'sufficiently accurate' and latency-free recognition is essential for general well being – if not for simply for survival in the real world- it is essential that biological agents are able to *adapt and generalize* known patterns quickly and effortlessly to recognize new variants that were not previously encountered. This plasticity is a necessary requirement for an organism's ability to adapt quickly and easily to new environments, fulfilling intrinsic *needs* (Maslow, 1954; Wang, 2003) as well as the capacity for dealing with previously unknown patterns of behavior and new varieties of patterns that have not been trained *explicitly*.

II. STATE-OF-THE-ART IN HUMAN SPEECH PROCESSING

Many branches of Cognitive Science, including (psycho-) linguistics, communication science, neurobiology, and brain imaging, have contributed to a large and rapidly growing mass of behavioral and phenomenological data about the speech processing skills of adult human beings and the ways in which these skills develop during infancy and childhood (Gerken & Aslin, 2005; Gopnik et al., 2001; Juczyk, 1999; Kuhl, 2004; Kuhl et al., 2003; Swingley, 2005). In accordance with the conventional cognitive approach, these data are almost invariably in the form of descriptions of the formal structure of utterances, rather than in terms of testable theories that can explain how the brain develops to exploit powerful and effective capabilities that enable humans to communicate with their environment. As a consequence, there is as yet no comprehensive and integrative theory that can even begin to explain ***how*** infants acquire speech and language (Werker & Yeung, 2005), nor ***how*** an adult's speech processing can be as fast and robust against all kinds of novel and adversary conditions as it apparently is. Today there is no computational model that can explain the acquisition of language and communication skills purely on the basis of sensory input, combined with the needs of infants to communicate in order to survive and flourish.

The Cross-channel Early Lexical Learning (CELL) model (Roy & Pentland, 2002) is very interesting in this aspect. CELL is a computational model for leaning associations and correlations (also known as 'words) in multimodal input data. It learns words by associating fragments in the speech input with the contents of pictures that contain objects that are referred to by the speech signal. The CELL model is interesting because it shows the potential of learning by cross-correlating asynchronous information across different input modalities. However, from the point of view of embodied cognitive systems, some aspects of the CELL model are less convincing. Most importantly, it is assumed that learning starts from what is in effect a discrete symbolic representation of the speech input, in the form of a lattice of symbols representing the phonetic sound sequence.

In contrast, there is growing evidence that speech and language skills are *emergent* capabilities of a developing communicative system (Johnson, 2002; MacWinney, 1998) and that the way in which linguistic patterns are stored and used during language acquisition and use change constantly as these patterns become more numerous and fine-grained, and as the methods needed for processing the patterns become correspondingly more complex (Werker & Curtis, 2005). In the same vein, there is growing evidence that 'processing' in many cases is based on direct links between perception and action, not mediated by conscious reasoning (Rizzolatti and Arbib, 1998).

In ACORNS, we opt for using the memory-prediction model of natural intelligence proposed by Hawkins (2004) to model what is perhaps the most complex cognitive capability in the known universe: speech processing. Drawing on the hierarchical structure of the neo-cortex, this theory proposes that a rich (and in terms of Shannon's information theory redundant) representation of speech input is first stored in a raw form in the lower levels of the neo-cortex. The stored representations comprise substantial detail, including the personal voice characteristics of the speaker, which is important for infants to be able to recognize their mother's voice. Of course, the capability to detect and process personal characteristics of voices remains important life long. Therefore, we must assume that people store more than just abstract symbolic representations of speech. Rather, the neural representation must comprise substantial additional detail. Which features exactly will be used in this representation, how emergent patters will be discovered and how abstraction takes place is topic of investigation in ACORNS.

It is known that on all levels of the cortical hierarchy perceptual patterns can be associated with (motor) actions. This enables various degrees of automation and shortcuts in *perception-action* loops, which are vital for latency-free responses in dangerous situations. The same mechanisms are instrumental in casual interactions with fellow human beings, where we show that we know that we are being addressed the moment an interlocutor starts talking and where we interpret the attitudinal (and emotional) layer of the speech well before the end of a turn (Thórisson, 2002). This is directly related with the shift from a data-driven bottom-up pattern recognition

towards a paradigm in which speech recognition is the result of bottom-up hypothesis construction and top-down verification. Once a pattern has formed in one of the higher layers in the cortical hierarchy, sensory input that is compatible with part of the pattern gives rise to top-down activation of the complete pattern. In other words: the brain 'predicts' that the sensory input will also contain the remaining parts of the pattern. And as long as the actual input does not deviate so much from the prediction that this hypothesis cannot be maintained, sensory input is assimilated into the pattern. Inputs that deviate too much from known patterns to be acceptable as new variants are stored as potentially new patterns, which may be associated with entities in the external world if they re-occur in similar conditions.

In this way the system can learn new patterns. Biological agents learn in a situation where there is always some form of feedback, implicit in the form of what happens in the environment, or explicit in the form of supportive or corrective actions of a caregiver. It is difficult to conceive of language acquisition without frequent and intensive interactions with a caregiver. The behavior of the caregiver in these interactions will tremendously enhance the reinforcement of existing and the creation of new patterns.

III. THE STRUCTURE OF ACORNS

In ACORNS, the scientific focus is on integrating five different interrelated aspects, viz. auditory front-end processing, pattern discovery procedures, memory access and organization, information discovery, and communication in what is essentially a simulation of an embodied agent.

Front-end processing

ACORNS focuses on the processing of auditory signals. Therefore, it is necessary to define and implement an auditory front-end processor, i.e., a module that converts acoustic signals into a rich internal representation that can be used for learning new patterns and for recognizing known patterns. There is accumulating evidence that this internal representation must account for features of the input signals in multiple simultaneous temporal resolutions, with a lower limit in the order of 0.5 to 2 ms, and an upper limit of about 250 ms (Hermansky, 1996). The representations must be suitable to characterize and process essentially all ecologically relevant sounds, from approaching footsteps and honking cars, doors opening or objects being displaced, to infant and adult directed speech. Since all these sounds can occur simultaneously, the representation must be suitable for the different sources to be modeled independently (Cooke and Ellis, 2001.

Pattern discovery

In conventional pattern recognition systems the patterns to be recognized, as well as the primitive elements from which complex pattern can be formed, are defined *a priori*. For example, in the conventional approach to speech recognition the patterns to be recognized are almost invariably words, while the primitives are related to the phonemes of the language (i.e., the speech sounds that distinguish between one word and another, such as *big* and *pig* in English). However, it is now generally agreed that the representation of words as sequences of phonemes like beads on a string is not adequate (Ostendorf, 1999) and there is accumulating evidence suggesting that speech patterns may be stored in the form of episodes spanning syllables or complete words, if not multiword expressions (Ernestus et al., 2002: Goldinger, 1998).

While conventional ASR systems sidestep the task of detecting what suitable basic units can be because these are pre-defined by the developer, infants must solve the problem somehow in language acquisition. It is likely that the repetitive character in speech, even when covering non-adjacent structures helps to distinguish word-like patterns (Newport et al., 2004).

Unlike in printed text, in normal speech, and even in infant-directed speech words are not separated by silences. Rather, words blend and merge at their boundaries. This makes it necessary for a 'newborn' speech acquisition system to discover patterns in the continuous input stream that correspond to meaningful speech events and eventually also to phoneme-like units. This task is probably simplified at least to some extent by the fact that infant-directed speech often consists of several repetitions of the same words and phrases (Snow & Ferguson, 1997).

Memory organization and access

Cognitive theories of memory in the agree that it is useful to distinguish at least three types of memory: a sensory store in which all information is captured only for a very short time (in the order of 2 seconds), a short-term memory (also called working memory) that holds representations of sensory inputs and serves as a executive module that is able to compare new sensory inputs to previously learned patterns that are retrieved from long-term memory. ACORNS aims at the development of suitable computational representations of these memories and the processing that takes place that can fit in an embodied form, inspired by the memory-prediction model. The eventual architecture will be able to support latency-free perception-action loops, as well as some forms of symbolic processing and reasoning for processing novel utterances. This architecture will be discussed in more detail below.

An important aspect of memory processes is how representations of novel patterns can form and be stored. In addition to storing a complete representation of the input signal, short-term memory must also be able to form and hold 'codes' derived from these complete representations. These codes will then be used to activate patterns that are already present in the long-term memory. Activated patterns must be compared to a more complete representation of the input signal in the working memory in accordance with the memory prediction model, by verifying the likelihood of activated patterns on the basis of the full signal representation.

Information discovery and integration

Storing patterns in memory is only useful if there are efficient and effective techniques for retrieving them. All available behavioral data strongly suggest that memory for speech and language is organized in an associative manner. Therefore, ACORNS will investigate three methods for addressing associative memories that approach the problem from slightly different angles. All these approaches hold the promise of scaling to problems as large and complex as natural language processing. The first approach takes its guidance from content addressable memories (CAM), especially the form of CAMs that handle fuzzy and incomplete codes for addressing the contents of the memory. The second approach borrows from Latent Semantic Analysis (e.g. Bellegarda, 2000), developed for document retrieval, but also proposed as a model for speech understanding (i.e., comprehending the semantic content of spoken utterances). The third approach is based on the assumption that similar patterns are related to each other, so that it is possible to quickly identify all patterns that resemble the input to be recognized. Gopnik et al. (2001) argue that an infant actively explores its environment and even conducts experiments (involving other people) to confirm/deny its developing capabilities.

Interaction and communication

Speech and language acquisition happen as the result of purposeful interaction between an infant and its environment. Therefore, it is essential to integrate all processing to realistically simulate speech acquisition driven by the intrinsic desire of an artificial agent to communicate with its environment. In the beginning an infant interacts with only a limited number of 'biological' agents. This will inevitably result in learning patterns that are strongly biased towards the personal voice characteristics of the caregivers. However, the infant will increasingly be addressed by other persons, thereby forcing the representations to generalize. From the very first days of its life, successful communication will contribute to fulfilling the most basic needs of the infant. However, in the case of an infant acquiring speech and communication skills it is difficult to map Maslow's hierarchy of needs onto its behavior, if only because Maslow's formulations address relatively abstract and high level needs. We assume that a newborn satisfies the physiological needs at the same time as the safety and love needs. For an artificial agent it is even more difficult to map Maslow's hierarchy (cf. Sarma & van der Hoek, 2004 for an attempt to adapt Maslow's hierarchy of needs to the situation where an individual is replaced by a team of software experts).
Therefore, in ACORNS it is proposed that the learning algorithm will be endowed with the intention to learn a continuously growing vocabulary in order to maximize the appreciation it receives from its environment. In the first stage of the project, ACORNS aims at being able to respond to 10 different words, a number that show grow later in the project.

IV. ARCHITECTURE

In ACORNS, we envisage to use the architecture depicted in figure 1. This architecture is based on recent psycholinguistic research on the organization of memory in connection to speech and language processing, especially on perceptual organization and its implication for short term memory (Jones, Hughes, and Macken, 2006).
During its speech interactions, the ACORNS system is presented two types of input from outside:

- multimodal data consisting of (a) explicit audio and (b) some kind of visual or tactile representation in of the object or concept expressed if the speech is infant-directed, while no systematic corresponding non-speech input will be provided with adult-directed speech.
- feedback about the appropriateness of output of the learning system, on an utterance by utterance basis. This information is equivalent to the feedback of the caregiver on the learning system's reaction to the input.

Multimodal input

The multimodal input consists of an audio stream (containing infant-directed speech, but also some adult-directed speech) in combination with an abstraction of the non-speech (visual/tactile) modality. This input is provided in synchrony with the speech. The audio stream is represented as sampled data streams; the non-speech modality contains an abstract representation that is associated to objects that are referred to in the infant-directed audio input. Thus, the learning system will not see the orthographic representation of the speech (in the form of invariant word symbols separated by blanks); instead, it listens to the audio input and has access to the *abstract* representation of the one (or two) subjects referred to by the speech signal. In this way, we simulate the presence of a visual sensory processing system of which the actual technical implementation and realization is not feasible given the time and resource constraints. It is up to the learning agent to learn word-like entities from the repetitions in the audio signal and from cross-modally reoccurring systematic patterning.

Memory architecture

The proposed architecture is based on results reported in recent literature on human memory, cognitive and language processing. It must be realized that a substantial part of this research is deigned and performed in a cognitive framework, focused on symbolic processing. However, we believe that the essential aspects of the cognitive model will also return in memory models of fully embodied systems. It is generally assumed that memory is divided into three types: sensory store, short-term memory (also known as working memory) and

long-term memory. The sensory store records ALL incoming sensory data. It is functionally partitioned into an 'echoic' memory (for acoustic information - lost after 2 sec) and an 'iconic' memory (for visual information - lost after 0.5 sec). To reduce computational load and to decrease the required amount of energy and time, an attention mechanism selects specific information from the sensory store and pushes this information into a short-term memory. This attention process actively performs a selection of the data in the sensory store to be stored in short-term memory.

The short-term/working memory can store data for up to about a minute. Its role is manifold:

- serve as a central execution platform
- perform dedicated tasks (e.g. for visual information it serves as a sketchpad, for speech it supports phone detection tasks)
- store (episodic) traces.

The storage capacity of the short-term/working memory is fixed; thus, new data overwrites older data.
The results available in short-term/working memory can be stored in long-term memory. This process (enhanced by rehearsal) is facilitated by the repetition of intrinsic and extrinsic presentations. Rehearsal of an *extrinsic* presentation may be forced by the frequent occurrence of a specific entity (e.g. a target word) in the input speech stream. Another way of rehearsal is *intrinsic*, in which the rehearsal is result of internal reflection on a certain representation. Data in long-term memory can remain a very long time, but may get lost due to interference. Also *access* to long-term memory can be lost to interference (see also http://thebrain.mcgill.ca)
The long-term memory is divided into two subparts: explicit memory and implicit memory.
The explicit (or declarative) memory is divided into episodic and semantic memory. The episodic store contains events and their contexts, and small samples of episodic traces. The semantic memory is assumed to contain the information that one is aware of, and that is usually independent of time and place, e.g. one's birthday, knowledge of the world, abstract relations and meaning. The implicit memory contains motor and process memory (necessary to e.g. play a piano or driving a car).
It is known that the content of a stored representation can change by the memorization process itself: the content of the memory is 'overlaid' by the context associated with the moment of memorization. Memory is thus a dynamically changing ensemble of representations with their contexts. Items in long-term memory can be accessed by retrieval, i.e. the storage of long-term memorized representations back into working (short-term) memory.

Attention and rehearsal processes

The attention and rehearsal mechanisms are processes that work on representations stored in memory, transforming stored representations into possibly more abstract representations. The precise relation between these processes is not exactly known – for example, some authors interpret their results as if attention is *maintained through rehearsal* in order for information to be stored in short-term memory. Attention is a process that reduces the part of the input stream that must be analyzed and is therefore indispensable for managing time, space, effort and in the end for being successful: to keep the computation load manageable, to reduce the storage into short-term (working) memory, to reduce the ambiguity to be resolved during the search, and to keep promising input features within the attention 'beam'.

Hierarchical Temporal Memories (HTM)

In ACORNS, HTMs (Hawkins, 2004) will be used to describe the function of the memory on a conceptual level. HTMs describe a memory structure with multiple connected levels: data arrive episodically at the bottom level, and find their way up towards more abstract representations at higher levels. At the same time, from the higher levels, top-down information goes back to the lower levels. The eventual interpretations and representations are the result of both bottom-up and top-down information in one common framework. If the incoming data are sufficiently supported by the top-down expectation activated by the bottom-up input, evaluation is shallow and fast (shallow verification on high level); otherwise evaluations are required on lower levels, requiring more time and effort. Attention is directly related to the sensitivity to inputs that are unexpected.

Feedback

Both training and test will be implemented as a control feedback loop, in which the reference signal from the caregiver ('environment') is represented via another channel (indicated in Figure 1 by 'ground-truth reaction from caregiver during training').
The role of feedback will be essential for the learning behavior of the ACORNS model. In ACORNS, we envisage two types of feedback: one is external and has the caregiver in the loop. The caregiver interprets the output of the model and provides feedback about her appreciation back to the learning model. This loop supports the optimization of appreciation received by the model from the environment. Besides this external loop, a number of internal feedback loops will be available as well. These loops take into account the performance of the various learning algorithms, such as the quality of a certain parse, the time it takes to perform a certain action, the amount of resources required to disambiguate a certain input. This means that learning takes place over two loops at the same time: one short cycle loop taking place several times per utterance, and one external loop that takes place on an utterance-by-utterance basis.

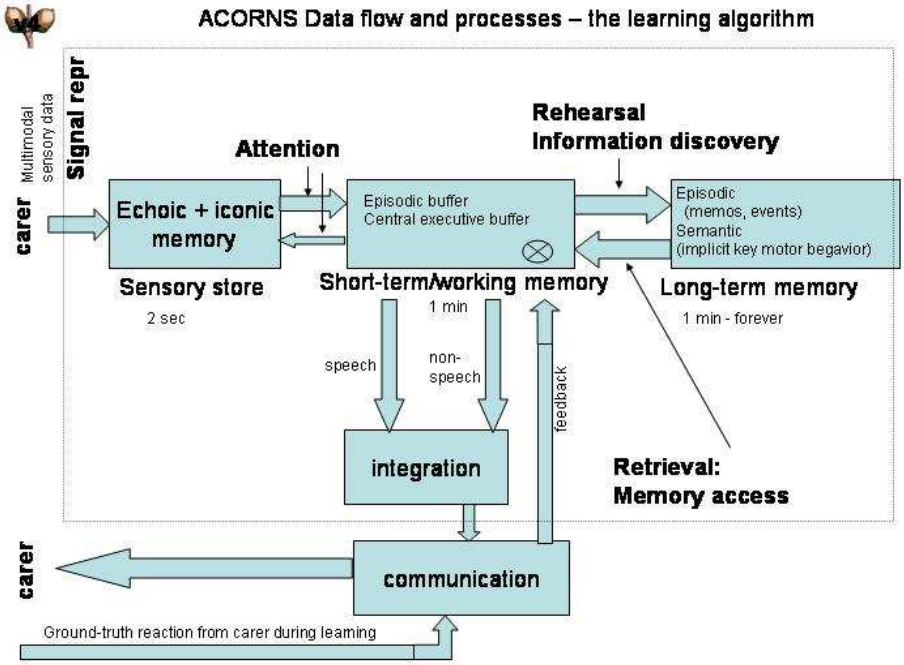

Fig.1 Global layout of the architecture. Multimodal input is presented to the model (upper left corner) and put into its sensory store. The three memory parts sensory store, short-term/working memory and long term memory form the entire memory, each with different decay times. Processes such as attention and rehearsal move stored representations from one type of memory to another in a possibly more abstract form. The integration module reads in multimodal abstract representations from the short-term/working memory and outputs the recognition result to the communication module. Two feedback loops are foreseen: one internal, governing the intrinsic learning processes and one external, in which the caregiver provides input to the model from outside based on the model output in the previous turn.

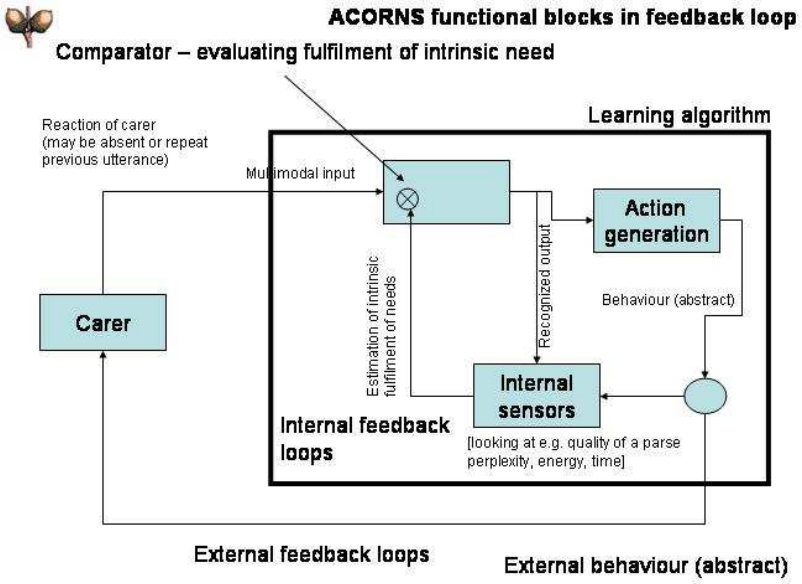

Fig. 2. Overview of the two types of feedback loops. In the external loop the caregiver interprets the output of the model and provides feedback on her appreciation on the model's output in the previous turn, followed by a new stimulus. This feedback takes place on an utterance by utterance basis. The internal feedback loops serve internal optimization processes and will run at a higher pace

V. ACORNS AND BEYOND

In ACORNS we take a realistic standpoint and we realize that we cannot investigate all aspects of the learning process in full detail. It is argued (Moore, 2007) that in order to understand the perception-action loop in detail the model must contain a speech production component which needs to learn and develop in the same loop as language comprehension. This production aspect has deliberately been taken out in the ACORNS plan, because it involves a technically potentially complicated aspect in the feedback loops. Also ACORNS refrains from modeling other kinds of active exploratory activities of the learning agent. It is interesting to speculate about the limitations that the lack of active exploration imposes on the agent and to investigate options for simulating language learning by an agent that is able to actively and purposefully explore its environment. To make experiments along this line feasible, it will probably be necessary to focus on an artificial agent that operates in virtual environments. Using physically embodied agents would result in experiments that take many years to perform. Moreover, the input that these 'robots' will receive is impossible to reproduce exactly.

ACORNS interprets prediction (as in the memory prediction framework) as a process involving both a recognition process and a memory structure in which future events can be anticipated. However, rolling out all possible future events, even in a dynamic way as in regular ASR, is unrealistic and undesirable. So, to reduce the computation load in short term/working memory we need an attention process. It is likely that this attention process is an emergent property in itself. The results form the ACORNS project will be useful in determining to what extent this attention process is itself a result of learning.

It is well known that the development of human auditory processing (and probably also of language acquisition) does not start at birth, but months before. It would be interesting to understand to what extent infants are tuned to rhythmic and prosodic patterns due to their exposure to pre-birth speech input (e.g. obtained by adequate low pass filtering). The idea of training the computational model of language acquisition would then be to feed the system with a speech signal with limited acoustic variation, to build abstractions on the basis of this impoverished input, and then to gradually enrich these abstractions by providing more and more details in the speech signal by opening up the bandwidth and the intrinsic variation in the signal.

Another limitation of the ACORNS project is that it will not be able to investigate the acquisition of syntactic structures with their attendant richer semantics that becomes available when full propositions can be processed. However, recent results in embodied construction grammar suggest that it will be possible to also apply the basic concepts of the memory-prediction theory to syntax and semantics (Feldman, 2006).

VI. CONCLUSION

The ACORNS project represents a first step towards establishing a computational model of the cognitive basis for the acquisition of spoken language. As it stands, research is primarily directed towards simulating the perceptual processes involved in the early learning of speech by young infants. It is envisaged that the insights gained through the investigation of HTM-based cortical mechanisms for modeling and predicting spoken language behavior will lead to substantial gains in terms of improving the understanding of the structure, representation and fidelity of speech signals. This, in turn will feed back into the design and construction of next-generation systems for automatic speech recognition, opening up applications hitherto forbidden by the fragile nature of current technology.

However, by excluding important components of the human system (such as speech production), the ACORNS architecture is fundamentally limited in what it is able to achieve. What is required for the future is an architecture for spoken language processing that extends ACORNS to a fully developed model that incorporates both perception and production in a balanced cognitive framework. Such an architecture has recently been proposed by one of the authors of this paper (Moore, 2007). Called PRESENCE – PREdicitive SENsorimotor Control and Emulation – this new model integrates findings from a range of scientific disciplines outside mainstream spoken language processing, and provides a unified architecture for modeling the interactive behavior of living organisms as well as a design for artificial cognitive systems. It is envisaged that PRESENCE will both influence the development of the ACORNS project, as well as evolve to incorporate the results of ACORNS research.

ACKNOWLEDGMENT

This research was funded in part by the European Commission, under contract number FP6-034362.

REFERENCES

Baddeley, A.D. (1986) Working Memory Clarendon Press, Oxford.

Bellegarda, J. R. (2000) Exploiting Latent Semantic Information for Statistical Language Modeling. Proc. IEEE, Vol. 88: 1279-1296..

Boves, L. & den Os, E. (1999) Applications of Speech Technology: Designing for Usability. Proc.IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, Co, 11-15.

Cooke, M. and Ellis, D. P.W. (2001). The auditory organization of speech and other sources in listeners and computational models, Speech Communication 35 (2001), pp. 141-177

Deerwester, S., Dumais, S. T., Furnas G. W., Landauer, T. K. & Harshman, R. (1990). 'Indexing by latent semantic analysis.' *Journal of the American Society for Information Science* 41, 391–407.

den Os, E.A., Boves, L., Rossignol, S., ten Bosch, L. and Vuurpijl, L. (2005) Conversational Agent or Direct Manipulation in Human-System Interaction. Speech Communication, 47: 194-207.

Ernestus, M., Baayen, R.H. & Schreuder, R.. (2002). The recognition of reduced word forms. *Brain and Language* 81, 162-173.
Feldman, J. (2006) From Molecule to Metaphor: A Neural Theory of Language, Cambridge, Mass: MIT Press.
Gerken, L., and Aslin, R.N. (2005) Thirty years of research in infant speech perception: the legacy of Peter Jusczyk. Language Learning and Development, 1: 5-21.
Goldinger, S. D. (1996) Words and voices: episodic traces in spoken word identification and recognition memory, J Exp Psychol Learn Mem Cogn, 22(5): 1166-1183.
Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review 105: 251-279
Gopnik A, Meltzoff A N, and Kuhl P K. (2001) The Scientist in the Crib, New York: William Morrow Co.
Hawkins, J. (2004) On Intelligence. New York: Times Books.
Hermansky, H. (1996) Auditory modeling in automatic recognition of speech. ESCA Workshop on the Auditory basis of speech perception, Keele University (UK), 15-19 July, 1996.
Hintzman, D. L.(1986) Schema-abstraction in a multiple-trace memory model, Psychological Review, 93: 411-427.
Johnson, S. (2002) Emergence. New York: Scribner.
Jones, D.M., Hughes, R.W. and Macken, W.J. (2006) Perceptual organization masquerading as phonological storage: Further support for a perceptual-gestural view of short-term memory, J. Memory and Language 54, 265 281.
Jusczyk, P.W. (1999) How infants begin to extract words from speech. TRENDS in Cognitive Science, 3: 323 328.
Kuhl, P.K. (2004) Early language acquisition: cracking the speech code. Nat. Rev. Neuroscience, 5: 831-843.
Kuhl, P.K. et al. (2003) Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. Proc. National Academy of Science U.S.A., 100: 9096-9101.
Lee, C.-H. (2004) From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Re-search Paradigm for Next Generation Automatic Speech Recognition. Proc. ICSLP.
Lippmann, R. (1997) Speech Recognition by Human and Machines. Speech Communication, 22: 1-14.
Maloof, M.A., Michalski, R.S. (2004). Incremental learning with partial instance memory. Artificial intelligence 154, 95-126.
Maslow, A. (1954) Motivation and Personality New York: Harper & Row.
Moore, E. and Clements, M. (2004), Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information, Proc. ICASSP 2004, May 14-17, Montreal, Canada
Moore R K. (2003) A comparison of the data requirements of automatic speech recognition systems and human listeners, Proc. EUROSPEECH'03, Geneva, pp. 2582-2584, 1-4.
Moore R K and Cunningham S P. (2005) Plasticity in systems for automatic speech recognition: a review, Proc. ISCA Workshop on 'Plasticity in Speech Perception, pp. 109-112, London, 15-17 June (2005).
Moore, R.K. (2007) Spoken Language Processing: Piecing Together the Puzzle. Accepted for Speech Communication.
Moore R K, Russell M J, Nowell P, Downey S N and Browning S R. (1994) A comparison of phoneme decision tree (PDT) and context adaptive phone (CAP) based approaches to vo-cabulary-independent speech recognition', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Adelaide.
Newport, Newport, E., Aslin, R. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. Cognitive Psychology, 48, 127-162.
Pfeifer, R. and Scheier, C. (1999) Understanding Intelligence. Cambridge, Mass.: MIT Press.
Ostendorf, M (1999) Moving beyond the 'beads-on-a-string' model of speech, in Proc. IEEE ASRU-99, Keystone, Colorado, USA. Dec 12-15.
Rizzolatti, G. & Arbib, M. A. (1998) Language within our grasp, Trends in Neuroscience 21, 188-194.
Roy, D.K. & Pentland, A.P. (2002) Learning words from sights and sounds: a computational model. Cognitive Science, 26: 113-146.
Sarma, A. and van der Hoek, A. (2004) A Needs Hierarchy for Teams. ISR Technical Report: UCI-ISR-04-9.
Shannon, C.E. & Weaver, W. (1949) The mathematical theory of communication. Urbana, Il.: University of Illinois Press.
Shalizi, C. R. and Crutchfield, J. P. (2000) Pattern Discovery and Computational Mechanics, Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2K).
Snow, C. and Ferguson, C. (1977). Talking to children: language input and acquisition. Cambridge, New York: Cambridge University Press.
Sroka, J. J. and Braida, L. D. (2005) Human and machine consonant recognition, Speech Communication: 44, 401– 423.
Swingley, D. (2005) Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology, 50: 86-132.
Thórisson, K. (2002) Natural turn-taking needs no manual: Computational theory and model from perception to action. In: B. Granström, D. House & I. Karlsson (Eds.) Multimodality in Language and speech systems. Dordrecht: Kluwer Academic, pp. 173-207.
Wang, Y. (2003) Cognitive Informatics: A new transdiciplinary research field. Brain and Mind, 4: 115-127.
Werker, J.F. and Curtis, S. (2005) PRIMIR: a developmental framework for of infant speech processing. Language Learning and Development, 1: 197-234.
Werker, J.F. and Yeung, H.H. (2005) Infant speech perception bootstraps word learning. TRENDS in Cognitive Science, 9: 519-527.
Wesker, T., Meyer, B., Wagener, K., Anemueller, J., Mertins, A. and Kollmeier, B. (2005) Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines. Proc. of Interspeech, Lisboa.