Chapter number 123456789

Discovery of words: Towards a computational model of language acquisition

Louis ten Bosch[a], Hugo Van hamme[b], Lou Boves[a]
[a]*Radboud University Nijmegen, the Netherlands*
[b]*Katholieke Universiteit Leuven, Belgium*

1. Introduction

Human speech recognition seems effortless, but so far it has been impossible to approach human performance by machines. Compared with human speech recognition (HSR), the error rates of state-of-the-art automatic speech recognition (ASR) systems are an order of magnitude larger (Lee, 2004; Moore, 2003; see also Scharenborg et al., 2005). This is true for many different speech recognition tasks in noise-free environments, but also (and especially) in noisy environments (Lippmann, 1997; Sroka & Braida, 2005; Wesker et al., 2005). The advantage for humans remains even in experiments that deprive humans from exploiting 'semantic knowledge' or 'knowledge of the world' that is not readily accessible for machines.

It is well known that there are several recognition tasks in which machines outperform humans, such as the recognition of license plates or barcodes. Speech differs from license plates and bar codes in many respects, all of which help to make speech recognition by humans a fundamentally different skill. Probably the most important difference is that bar codes have been designed on purpose with machine recognition in mind, while speech as a medium for human-human communication has evolved over many millennia. Linguists have designed powerful tools for analyzing and describing speech, but we hardly begin to understand how humans process speech. Recent research suggests that conventional linguistic frameworks, which represent speech as a sequence of sounds, which in their turn can be represented by discrete symbols, fail to capture essential aspects of speech signals and, perhaps more importantly, of the neural processes involved in human speech understanding. All existing ASR systems are tributable to the beads-on-a-string representation (Ostendorf, 1999) invented by linguistics. But is quite possible –and some would say quite likely- that human speech understanding in not based on neural processes that map dynamically changing signals onto sequences of discrete symbols. Rather, it may well be that infants develop very different representations of speech during their language acquisition process. Language acquisition is a side effect of purposeful interaction between infants and their environment: infants learn to understand and respond to speech because it

helps to fulfil a set of basic goals (Maslow, 1954; Wang, 2003). An extremely important need is being able to adapt to new situations (speakers, acoustic environments, words, etc.) Pattern recognisers, on the other hand, do not aim at the optimisation of 'purposeful interaction'. They are trained to recognize pre-defined patterns, and decode an input signal in terms of a sequence of these patterns. As a consequence, automatic speech recognisers have serious problems with generalisations. Although modern ASR systems can adapt to new situations, this capability is limited to a predefined set of transformations (Moore & Cunningham, 2005).

Can the gap in speech recognition performance between humans and machines be closed? Many ASR scientists believe that today's statistical pattern recognisers are not capable of doing this (see e.g. Moore, 2003). Most probably ASR can only be improved fundamentally if entirely new approaches are developed (Bourlard et al, 1996). We are trying to do just this, by investigating the way how infants acquire language and learn words and to see to what extent this learning process can be simulated by a computational model. Many branches of Cognitive Science, such as Psycho-linguistics, and Communication Science have contributed to a large mass of data about the speech processing skills of adults and the ways in which these skills develop during infancy and childhood (MacWhinney, 1998; Gerken & Aslin, 2005; Gopnik et al., 2001; Jusczyk, 1999; Kuhl, 2004; Kuhl et al., 2003; Swingley, 2005; Smith & Yu, 2007). Despite the large number of studies, it is not yet clear how exactly infants acquire speech and language (Werker & Yeung, 2005), and how an adult's speech processing can be as fast and robust against novel and adverse conditions as it apparently is. The design and use of a computational model is instrumental in pinpointing the weak and strong parts in a theory. In the domain of cognition, this is evidenced by the emergence of new research areas such as Computational Cognition and Cognitive Informatics (e.g. Wang et al, 2007).

In this chapter, we describe research into the process of language acquisition and speech recognition by using a computational model. The input for this model is similar to what infants experience: auditory and visual stimuli from a carer grounded in a scene. The input of the model therefore comprises multimodal stimuli, each stimulus consisting of a speech fragment in combination with visual information. Unlike in a conventional setting for training an ASR system, the words and their phonetic representation are not known in advance: they must be discovered and adapted during the training.

In section 2, we will present the model in more detail. The communication between the learner model and the environment is discussed in section 3. In section 4, the mathematical details of one specific instantiation of the learning algorithm are explained, while section 5 describes three experiments with this particular algorithm. Discussion and conclusion are presented in sections 6 and 7.

2. The model

2. 1 Background

In order to be able to effectively communicate, infants must learn to understand speech spoken in their environment. They must learn that auditory stimuli such as stretches of speech are not arbitrary sounds, but instead are reoccurring patterns associated with objects and events in the environment. Normally this development process results in neural representations of what linguists call 'words'. This word discovery process is particularly interesting since infants start without any lexical knowledge and the speech signal does not contain clear acoustic cues for boundaries between words. The conventional interpretation is that infants must 'crack' the speech code (Snow & Ferguson, 1977; Kuhl, 2004) and that the discovery of word-like entities is the first step towards more complex linguistic analyses (Saffran and Wilson, 2003). However, it seems equally valid to say that infants must construct their individual speech code, a complex task in which attention, cognitive constraints, social and pragmatic factors (and probably many more) all play a pivotal role.

Psycholinguistic research shows that infants start with learning prosodic patterns, which are mainly characterised by their pitch contours and rhythm. A few months later, infants can discriminate finer details, such as differences between vowels and consonants (e.g. Jusczyk, 1999; Gopnik et al., 2001). At an age of about 7 months infants can perform tasks that are similar to word segmentation (e.g. Werker et al., 2005 and references therein; Newport, 2006; Saffran et al., 1996; Aslin et al., 1998; Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). These skills can be accounted for by computational strategies that use statistical co-occurrence of sound sequences as a cue for word boundaries. Other experiments suggest that the discovery of meaningful 'words' is facilitated when the input is multimodal (e.g. speech plus vision), experiments (Prince & Hollich, 2005) and computational models (such as the CELL model, Roy & Pentland, 2002).

As observed above, the design and test of a computational model of word discovery may be pivotal for our understanding of language acquisition in detail. Simultaneously, such a model will inform possible ways to fundamentally alter (and hopefully improve) the conventional training-test paradigm in current ASR research. The classical limitations for defining and modelling words and phonemes in ASR might be radically reduced by exploring alternatives for data-driven word learning (e.g. by the use of episodic models – see Goldinger, 1998; Moore, 2003).

The computational model that we are developing differs from most existing psycho-liguistic models. Psycho-linguistic models of human speech processing (e.g. TRACE, McLelland & Elman, 1986; Shortlist, Norris, 1994; Luce & Lyons, 1998; Goldinger, 1998; Scharenborg et al., 2005; Pisoni & Levi, 2007; Gaskell, 2007) use a predefined lexicon and take symbolic representations of the speech as their input. The fact that a lexicon must be specified means that these models are not directly applicable for explaining word discovery (nor other aspects of language acquisition). The success of these models, however, suggests that concepts such as activation, competition and dynamic search for pattern sequences are

essential ingredients for any model aiming at the simulation of human speech processing (cf. Pitt et al, 2002, for a discussion about these topics).

The computational framework that we propose in this paper builds on Boves et al. (2007) and combines the concepts of competition and dynamic sequence decoding. Simultaneously, it builds the lexicon in a dynamic way, starting empty at the beginning of a training run. During training, the model receives new utterances, and depending on the internal need to do so, new representations are hypothesized if existing representations fail to explain the input in sufficient detail.

The model hypothesizes that patterns are stored in memory mainly on the basis of bottom-up processing. Bottom-up models performing pattern discovery are also described in Park & Glass (2006) and ten Bosch & Cranen (2007). These models are based on a multi-stage approach in which first a segmentation of the speech signal is carried out, after which a clustering step assigns labels to each of the segments. In the final stage, then, these symbolic representations are used to search for words. The important difference beween ten Bosch & Cranen (2007) on the one hand and Park & Glass (2006) and Roy & Pentland (2002) on the other is that former does not rely on the availability of a phonetic recogniser to transcribe speech fragments in terms of phone sequences. Models that do bottom-up segmentation have already been designed in the nineties by Michiel Bacchiani, Mari Ostendorf and others. But the aim of these models was entirely different from ours: the automatic improvement of the transcription of words in the lexicon (Bacchiani et al., 1999).

2.2 Architecture

Our novel computational model of language acquisition and speech processing consists of two interacting sub-models: (1) the carer and (2) the learner. In this paper we focus on the architecture of the learner model. The computational model of the learner must be able to perform three major subtasks.

Feature extraction
The learner model has multimodal stimuli as input. Of course, the speech signal lives in the auditiry modality. To process the audio input, the model has an auditory front-end processor, i.e., a module that converts acoustic signals into an internal representation that can be used for learning new patterns and for decoding in terms of known patterns. The front-end generates a redundant representation that comprises all features that have been shown to affect speech recognition (and production) in phonetic and psycholinguistic experiments. However, for the experiments described in this chapter we only used conventional Mel Frequency Cepstral Coefficients (with c_0) and log energy.
In the second modality (vision), we sidestep issues in visual processing by simulating the perception of objects and events in the scene by means of symbols (in the simplest version) or possibly ambiguous feature vectors (in more complex versions of the model).

Pattern discovery
The learning paradigm of the computational model is different from conventional automatic speech recognition approaches. In conventional speech recognition systems the patterns to be recognised are almost invariably lexical entries (words), representeds in the form of

sequences of phonemes. In the current model, we avoid the a priori use of subword units and other segmental models to hypothesize larger units such as words, and explicitly leave open the possibility that the model store patterns in the form similar to episodes (see also McQueen, 2007).

Apart from the question how meaningful (word-like) units can be *represented*, the discovery of words from the speech signal is not straightforward. In our model, we use two strategies: (1) exploit the repetitive character of infant-directed speech (Thiessen et al., 2005) (2) make use of the cross-modal associations in the speech and the vision modality. This is based on the fact that infants learn to associate auditory forms and visual input by the fact that the same or similar patterns reappear in the acoustic input whenever the corresponding visual scene is similar (Smith & Yu, 2007; see also Shi et al, 2008).
The chosen architecture is such that representations of word-like units develop over time, and become more detailed and specialised as more representations must be discriminated.

Memory access
Theorists on the organisation of human memory disagree on the functioning of human memory and how exactly the cognitive processes should be described. However, there is consensus about three processes that each plays a different role in cognition (MacWhinney, 1998). Broadly speaking, a sensory store holds sensory data for a very short time (few seconds), a short-term memory (holding data for about one minute) acts as 'scratch pad' and is also used for executive tasks, while a long-term memory is used to store patterns (facts e.g. names and birthdays, but also skills such as biking) for a very long time.

The short-term memory allows to store a representation of the incoming signal (from the sensory store) and to compare this representation to the learned representations retrieved from long-term memory. If the newly received and previously stored patterns differ mildly, stored representations can be adapted. If the discrepancy is large, novel patterns are hypothesized and their activation is increased if they appear to be useful in following interactions. If they are not useful, their activation will decay and eventually they will not be longer accessible. Short-term memory evaluates and contains activations, while long-term memory stores representations.

The input and architecture of the computational model are as much as possible motivated by cognitive plausibility. The words, their position in the utterance, and its acoustic/phonetic representation are unspecified, and it is up to the model to (statistically) determine the association between the word-like speech fragment and the referent.

3 Interaction and communication

Language acquisition takes place in communication loops between the infant and the environment. In the beginning of language acquisition, the number of persons that the infant interacts with is usually limited, which leads to patterns that are biased towards the personal voice characteristics of these few caretakers. As soon as the infant is addressed by more persons, the stored representations will be adapted in some way to accommodate the differences between speakers (and other differences, such as speaking styles).

The communicative framework involves two active participants and simulates a 'learner' involved in interaction with a 'carer'. The learner discovers words and word-like entities on the basis of the grounded stimuli presented by the carer during the interaction.
The learner starts with an almost empty memory, and during the interaction between learner and carer, the learner gradually detects more and different meaningful sound patterns. This is done by first hypothesising an internal representation of a word-like entity, followed by strengthening or weakening of this representation on the basis of new stimuli. This means that the concept of word is not built-in a priori, but that meaningful acoustic patterns come about as an emergent property during learning. 'Words' in the linguistic sense of the term are meta-level concepts that children acquire when they start talking *about* language.

The multimodal stimuli from which our model must learn consist of two parts (a) the audio parting the form of real speech signals (short utterances) and (b) the visual (semantic) input corresponding to the meaning of the utterances. This visual representation in the experiments described here is an abstract tag, which uniquely refers to the object that is referred to by the utterance. In the experiments described in section 5, we use 13 of these tags (representing 13 target words). The tags represent an abstraction of the information that would otherwise be available along the visual modality. The tag itself does not give any clue about the word, or the phonetic representation of any target word.

The speech used for training the learner is highly repetitive in terms of verbal content and produced by four speakers. In a later phase of the learning process the model will be exposed to speech produced by other speakers. The communication starts when the carer presents a multimodal stimulus to the learner. Once the carer has provided a stimulus, the learner's response consists of the concept the learner thinks is referred to by the audio part of the stimulus. This reply is combined with a confidence measure. In the learner's memory this results in an update of the internal representations of the concepts.
The emphasis is on learning a small vocabulary, starting with an empty lexicon. A basic vocabulary must be formed by listening to simple speech utterances that will be presented in the context of the corresponding concepts.

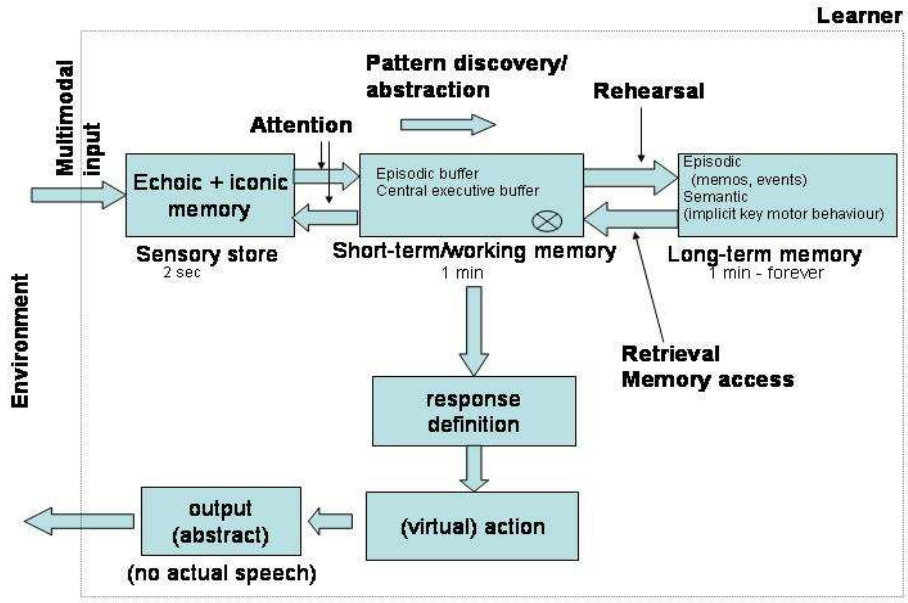

Figure 1. This picture shows an overview of the overall interaction between learner model (within grey-line box) and the environment (i.e. carer, outside the box). Multimodal stimuli are input of the model (top-left corner). For an explanation see the text.

A schematic representation of the interaction between the learner and the carer is shown in figure 1. The learner is depicted within the grey box, while the carer is indicated as the environment outside the grey box. A training session consists of a number of interaction cycles, each cycle consisting of several turns. Per cycle, the learner receives multimodal input from the carer after which a reply is returned to the carer. In the next turn, the carer provides the learner with a feedback about the correctness of the response, after which it is up to the learner to use this feedback information.
When the learner perceives input, the speech input is processed by the feature extraction module. The outcome is stored in the sensory store, from where it is transferred to short-term memory (STM) if the acoustic input is sufficiently speech-like (to be determined by the attention mechanism in Figure 1). In STM, a comparison takes place between the sensory input on the one hand and the stored representations on the other. The best matching representation (if any) is then replied to the carer.

The role of the carer
The carer provides multimodal utterances to the learner. The moment at which the carer speaks to the learner is determined by a messaging protocol that effectively controls the interaction during a training session. The utterances used during training and their ordering

are determined by this protocol. After a reply from the learner, the carer provides feedback about the correctness of the reply. In the current implementation, the feedback is just a binary yes/no (approval/disproval).

Learning drive

The communication between carer and learner is not enough for learning. Learning is a result of a learning drive. Exactly which drive makes the learner learn? When looking at real life situations, a baby's drive to learn words is ultimately rooted in the desire to have the basic needs for survival fulfilled: get food, care and attention from the carers. In the current model, this 'need' is implemented in the form of an 'internal' drive to build an efficient representation of the multimodal sensory input, in combination with an 'external' drive to optimise the perceived appreciation by the carer.

The internal drive basically boils down to the quality of the parse of the input. Given a certain set of internal representations, the leaner is able to parse the input to a certain extent. If the input cannot be parsed, this means that representations must be updated or even that a new representation must be hypothesised and stored.

The external drive (related to the optimisation of the appreciation by the carer) is directly reflected in the optimisation of the accuracy of the learner's responses (i.e. minimisation of the error rates). The optimisation of the accuracy can mathematically be expressed in terms of constraints on the minimisation between predicted reply (predicted by the learner model) and the observed ground truth as provided in the stimulus tag.

4 Learning and decoding algorithm

In the current implementation of the learner's model, training and decoding is done by first combining acoustic and visual/semantic information from the incoming new stimulus into one single vector. Thus, each stimulus is represented as a vector with a fixed dimension. As a result, a sequence of stimuli is represented as a matrix (in this chapter, this data matrix will be referred to by X). The actual search for patterns is performed by a decomposition technique called Non-Negative Matrix Factorisation (NMF) on X (Hoyer, 2004; Stouten et al, 2007, 2008). NMF is a technique to find structure in (large) data sets. The usefulness of NMF for our purpose derives from the fact that it is able to decompose the very large matrix X into two (much smaller) matrices W and H such that

(a) $X \approx WH$: The distance between X and the product WH is minimised according to some distance function (see below)
(b) All components of X, W and H are positive or zero

In our experiments, X represents previously learnd (but constantly updatable) patterns residing in the long term memory of the learner. Prior to the training X is initialised to the empty matrix. Each new utterance is first encoded in a vector which is then appended to the current matrix X. If there is no forgetting, the number of columns of X equals the number of utterances observed so far in the training run. This is reminiscent of episodic memory.

After NMF decomposition, the columns in W act as basis vectors into which the columns of X are represented. What we want the NMF to produce is a decomposition of each utterance

in terms of word-like entities. In the current experiments, where X is based on utterances (in linguistic term: sequences of words), each column of W should therefore ideally be related to a single word-like entity. A necessary condition to enable this utterance–to-word decomposition is provided by the way in which new utterances are mapped to a new column of X. This mapping will be denoted *map* below. If s_1 and s_2 are two arbitrary speech segments, and $s_1 \oplus s_2$ is the concatenation of these segments, then *map* must meet the following property:

$$map(s1 \oplus s2) = map(s1) + map(s2)$$

where the right-hand side '+' sign refers to the usual vector-addition of two vectors.

The matrix H contains the corresponding weights (activations) in the decompositions. If a certain entry in H (h_{ij}) is positive, it means that the corresponding column of W is required with weight h_{ij} to explain the corresponding column in X.
In this decomposition, the number of columns in W (and so the number of rows in H) is a model parameter. This number must be estimated on the basis of the number of input stimuli observed so far. In the current implementation, this number is initialised to 5 and increases with the number of observed stimuli. The way how this number increases is basically heuristically determined. The number of columns in W must be larger than the number of concepts that must be distinguished. Else, it would not be possible to account for different linguistic contexts of the target words (the acoustic context of the target words in the utterances). This implies a tendency for the learner to overestimate the number of things in the environment that must be distinguished.
In the current implementation, the NMF factorisation takes place for the first time after K stimuli have been observed, where K is a user-defined number. This corresponds to the assumption that reorganization towards a more efficient storage of episodic traces in the memory is only necessary after a certain number of observations. After this initialisation, the resulting W and H are updated after each following stimulus. As a result, W and H evolve gradually as long as more utterances are being observed.

To *decode* a new (not yet observed) utterance U, the *map* operation is applied on U, and a vector h is sought such that the difference between

$$map(U) \text{ and } W\,h$$

with W the *current* internal representation, is minimised. As a result, the vector h encodes the utterance in terms of activations of the columns of W: The winning column is the one corresponding to the highest value in h.
As said above, the multimodal stimulus contains a tag corresponding to visual input; this tag can be coded into W such that each column of W is statistically associated with a tag. In combination with the information in the vector h, this association allows the learner to respond with the corresponding tag, in combination with the corresponding value in h.

NMF minimisation

The minimisation of the NMF cost function leads to the overall closest match between prediction and observation, and so to an overall minimisation of the recognition errors made by the learner. Hoyer (2004) presents two different NMF algorithms, each related to a particular distance that is to be minimised. In the case of minimisation of the Euclidean distance (Frobenius norm) between X and WH, the cost function that is minimised reads (see Hoyer, 2004 for details)

$$F_1(X, WH) = \sum_{i,j} (X_{ij} - [WH]_{ij})^2 / 2$$

while in case of the Kullback-Leibler divergence, this cost function reads

$$F_2(X, WH) = \sum_{ij} (X .* \log(X ./ (WH) - X + WH)_{ij}$$

In this formula, .* and ./ denote component-wise multiplication and division, respectively. The structure of the expressions at the right-hand side indicates that in both cases the error between prediction and observation is an accumulated sum over all tokens that are available in X (and so processed during training). Splitting the 2-way sum in two separate one-way sums (using i and j, respectively), the terms $(X_j\text{-}[WH]_j)^2$ and $(X .* \log(X./WH) - X + WH)_j$ can be interpreted as the internal target function that is to be minimized on a token-by-token basis. In these token-related expressions, X, WH and H are now column vectors rather than matrices; W is a matrix.

The second expression, related to the Kullback-Leibler distance between reference (X) and hypothesis (WH), can be regarded as the log-likelihood of the model (XH) predicting the observation (X). This, in turn, can be interpreted as a measure for the quality of the parse of the utterance associated to X, in terms of the concepts that are associated with the columns in matrix W.

Interestingly, the learning model does not need to segment utterances in order to hypothesize the presence of target words. Nor is the ordering of the words in the utterance used. This is so because the *map* function is symmetric:

map(s1 ⊕ s2) = *map*(s1) + *map*(s2) = *map*(s2 ⊕ s1)

This implies that word ordering is not reflected after the map operation. We come back to this property in the discussion section (section 6).

Figure 2 shows how the use of NMF relates to the use of abstraction, which is necessary to facilitate access to the internal representations of speech signals. It shows how information

in X (at a low level of abstraction) is factorised to obtain more abstract information (in W and H). Thus, abstraction is mathematically modelled as factorisation of the data matrix X.

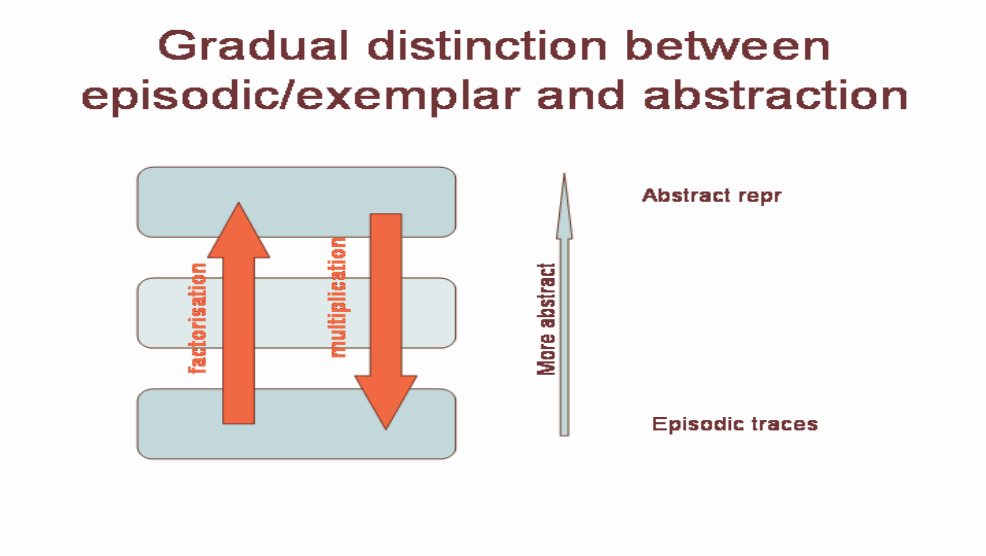

Figure 2. This figure shows a multi-layered representation of the contents of the memory of the learner model. On the lowest level, data are represented in unreduced form. The higher the level, the more abstract the corresponding representation is. The picture shows the general idea of having multiple different levels of abstractness in parallel. In the current computational model of the learner, just two levels are used, an 'episodic' one (here: actual sequences of feature vectors obtained from the feature extraction module), and an abstract one (here: basis vectors in a vector space representing words, in combination with activation strengths). By using NMF, the conceptual bottom-up process of abstraction is translated into explicit matrix factorisations, while the top-down process is represented by matrix multiplications. These top-down and bottom-up processes can interact in a natural way since they use the same paradigm of algebraic matrix manipulation.

5 Experiments

5.1 Materials

For the experiments discussed here, we use three comparable databases collected in the ACORNS project: one Dutch database (NL), a Finnish database (FIN), and a Swedish database (SW). For each language, the databases contain utterances from 2 male and 2 female speakers. Each speaker produced 1000 utterances in two speech modes (adult-directed, ADS, and infant-directed, IDS). For the infant-directed style, all speakers were asked to act as if they addressed a child of about 8-12 months old. The resulting speech has the well-known characteristics of infant-directed speech, such as a more exaggerated intonation, clear pronunciation, and low speaking rate.
The set of 1000 uterances contains 10 repetitions of combinations of target words and 10 carrier sentences. Within a database, not all target words are uniformly distributed. While all 4 speakers share the same target words, the proper name they use to address the learner is different for each speaker. For example, the NL database (8000 utterances) contains 800 tokens of ecologically relevant target words such as luier (diaper), auto (car), but only 200 of the proper names mirjam, isabel, damian, otto. In total, there are 13 different target words per language.

5.2 Experimental set-up

In each experiment the training is based on a specific list of utterances selected from the available pool of 8000 utterances. The ordering in which the utterances are presented is one of the experimental parameters (for example, this ordering can be random or speaker-blocked). During a training the utterances are always processed utterance-by-utterance. It must be emphasized that in our approach there is no essential difference between training and test: each NMF update is based on the history observed so far (matrix X), while each new utterance is recognised (decoded) on the basis the stored representations (W) of the learner learned so far. In the experiments reported here, the length of the history-update window used in each NMF step is a parameter. One training session of the computational model consists in presenting (by the carer) the next not yet observed multimodal stimulus. The learner attempts to decode the audio part of this new input stimulus, and replies by providing its most active word hypothesis in combination with a confidence score. Then, exactly as in a real-life carer-child interaction, it is up to the carer-model to give feedback: by providing the next stimulus, or by correcting the model's reply.

5.3 Experiment 1

Experiment 1 aims at showing that the learner is able to create representations of target words, and that when a new speaker is encountered, these representations must be adapted towards the characteristics of the new speaker.

To that end, the pool of 8000 Dutch utterances was blocked by speaker, and randomized within speaker. The resulting utterance list contained 2000 utterances by a female speaker,

followed by 2000 utterances produced by a male speaker, followed again by the utterances from another female and another male speaker.

The results of this word detection experiment are shown in Figure 3. The plot shows the performance of the learner, measured as average accuracy over the most recent 50 stimuli. The horizontal axis shows the number of stimuli (tokens) presented so far. The vertical axis shows the corresponding accuracy in terms of percentages correct responses. Each time a new speaker starts, a drop in performance of about 20-30 percent points can be seen. This performance drop is mainly due to the fact that the word representations learned so far are inadequate to correctly parse the utterances by the new speaker. The dip shows that representations are dependent on the speakers previously encountered during training.
Given the learning settings, the learner is able to create adequate internal representations for 10 target words as produced by the first female speaker within about 1000 tokens (that is, approximately 100 tokens per word). For each new speaker, the performance is back on its previous high level within about 100 tokens per word. Results for Finnish and Swedish are very similar.
During the first few hundred utterances the learner does not have any repreentatin available and so does not respond in a meaningful manner; this explaines why the accuracy is zero.

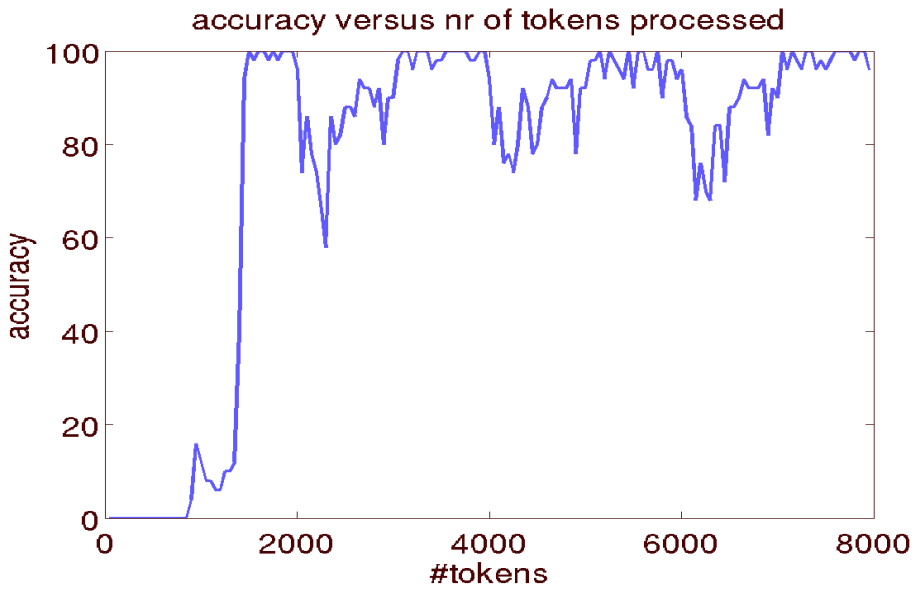

Fig 3. Results of word detection experiment (for Dutch, speaker-blocked). The plot shows the performance of the learner, measured as average accuracy over the most recent 50 stimuli. The horizontal axis shows the number of stimuli (tokens) presented so far. The vertical axis shows the corresponding accuracy in terms of percentages. A drop in performance of about 20-30 percent point can be seen each time when a new speaker starts.

5.4 Experiment 2

During the training, the NMF update takes place after each utterance. Tus, there are two parameters in the model that affect the eventual performance of the learner. These parameters specify the update scheme for the internal representations: how many utterances are to be used in the update of the internal representations, and when the initialisation of the internal representation should occur. The first parameter (number of utterances used in each NMF step) is referred to by memory length (indicated by 'ml') – this parameter specifies something that might be called 'effective memory length'. The second parameter deals with the initialisation and denotes the number of stimuli before the first NMF decompistion ('nsbt').

In this experiment, we focus on the 2000 utterances of one Dutch female speaker. Figure 4a shows the dependency of the eventual performance of the memory length. Four values for ml are shown (20, 100, 500, inf). The value 'inf' means that all utterances that are observed so far are used in the NMF updates. In this experiment, the value of nsbt is fixed to 100, which means that the very first NMF factorisation occurs after 100 utterances, after which recognition takes place.
The plot shows that the eventual performance largely depends on the memory length. Values of 500 and 'inf' do lead to results that are almost indistinguishable; a value of 100, however, leads to considerably lower performance. Translating this to the level of individual words, this implies that 50 tokens per word suffice, but 9 to 10 tokens are insufficient to yield adequate representations.
As shown in Fig. 4b the effect of the parameter nbst is much less dramatic. The most interesting observation is that there is no need to delay the first decomposition until after a large number of input stimuli have been observed. Delaying the first decomposition does not buy improvements in later learning. But in a real learning situation it might cost a baby dearly, because the carer might become frustrated by the lack of meaningful responses.

5.5 Experiment 3

In this experiment, the aim is to show that internal representations are changing continuously, and that we can exploit structure in the representation space by statistical means. This shows how abstraction may follow as a result of competition in crowded collections of representations on a lower level. For example, we would like to know whether speaker-dependent word representations can be grouped in such a way that the common characteristics of these representations combine into one higher-level word representation. We investigate this by first creating speaker-dependent word representations, followed by a clustering to arrive at speaker-*in*dependent word representations.

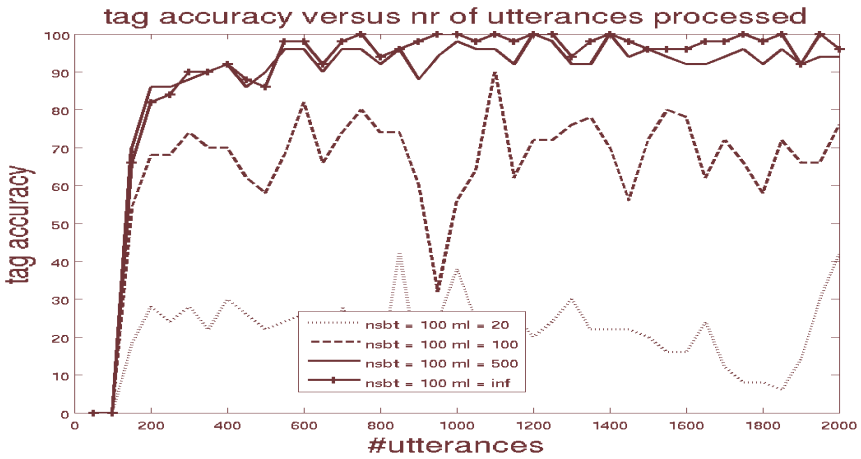

Figure 4a. This figure shows the dependency of the eventual performance of the memory length. Three values for memory length (indicated by 'ml') are shown (20, 100, 500, inf). The value 'inf' means that all utterances that are observed so far are used in each NMF update. The number of stimuli that are processed before the first NMF-step ('nsbt') is fixed to 100. For further explanation see the text.

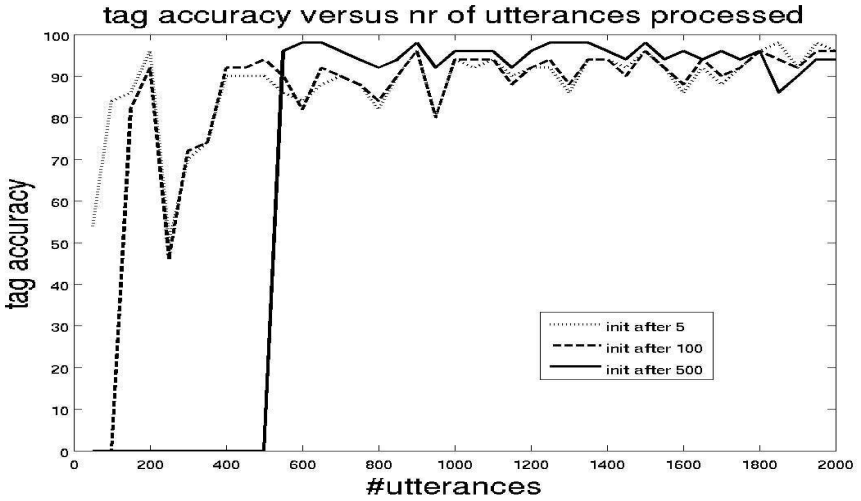

Figure 4b. In this figure, the performance of the learner is shows as a function of the number of utterances used for the first NMF update ('init'). For the sake of comparison, the memory length is chosen to be equal to 500. The dashed curve in this figure is comparable to the solid curve in figure 4a (ml = 500, number of stimuli used in first NMF factorisation = 100). One observes that the eventual learner result is only slightly dependent on the amount of data used in the initialisation of W and H.

The training data are taken from the Dutch database and consists of 2000 utterances, 500 utterances randomly chosen from each speaker. The visual tags that are associated to the utterances now differ from the tags used in the two previous experiments. While in those experiments the tag was a unique reference to an object, such as 'ball', the tags in this experiment are a combination of the object referred to (ball) *and* the speaker. That means that the leaner has to create and distinguish speaker-dependent representations for all 'words', leading to 36 different columns in the W matrix (the nine common words x four different speakers). As a result, each column encodes a speaker-dependent variant of a target word. For example, for the single target word 'luier' (diaper), 4 columns in W represent the speaker-dependent acoustic realisations as produced by the four speakers.
The question in this experiment is to what extent the W columns can be clustered such that the speaker-dependent variants of a single word can be interpreted as belonging to one cluster.
All representations are one-to-one with columns in W. The metric of the vector space in which these columns reside is defined by the symmetrised Kullback-Leibler divergence. This means that for any vector pair (v_1, v_2) the distance KL(v_2, v_2) can be used as a dissimilarity measure, resulting in a KL-distance matrix M_{KL}. A 10-means clustering using M_{KL} then yields 10 clusters (where each cluster contains one or more word-speaker representations).
Eventually, we obtained clusters that correspond almost perfectly to speaker-independent word representations. Figure 5 shows how the between-cluster distance increases while the average within-cluster variance decreases during training. This implies that clusters do emerge from the entire set of representations, which indicates that NMF is able to group speaker-dependent word representations one more abstract representation.

One interesting aspect to address here is the precise evaluation of the within and between-cluster variances. This is not trivial, since the KL divergence in the vector space spanned by the columns of W is not Euclidean, meaning that the concept of 'mean' vector is problematic. To circumvent this, the symmetrised KL divergence was first used to define a distance between any two vectors in the space spanned by the columns of W. Next, evaluation of the mean vector was avoided by making use of the following property:

$$\sum_i (x_i - < x >)(x_i - < x >)^t = 0.5 \sum_{ij} (x_i - x_j)(x_i - x_j)^t$$

Application of this expression for both within and between cluster variances leads to the results as shown in Figure 5.

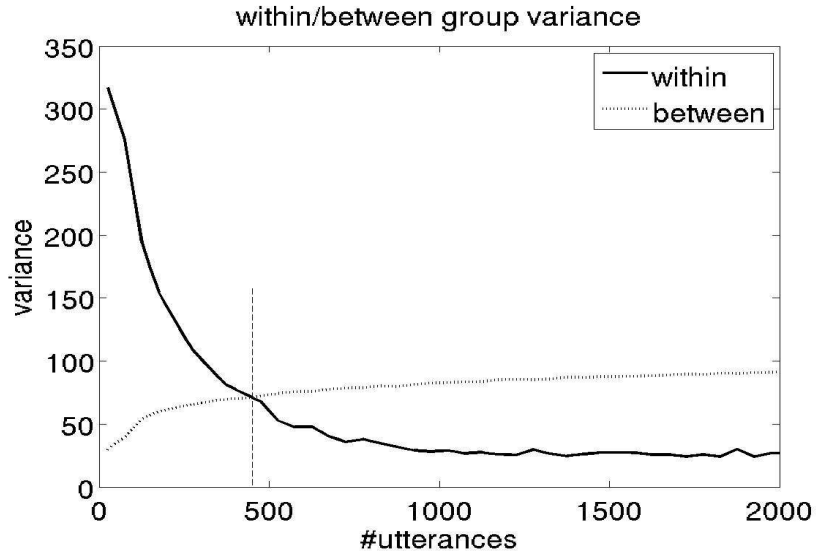

Figure 5. Values of the between-cluster variance (dotted line) and within-cluster variance (bold line) during training. The ratio of the within-variance and between-variance decreases. This shows that the speaker-dependent word representations can indeed be clustered into groups that become increasingly more distinct.

6 Discussion

The computational model presented here shows that words (and word-like entities) can be discovered without the need for a lexicon that is already populated. This discovery mechanism uses two very general learning principles that also play a role in language acquisition: the repetitive character of infant-directed speech on the one hand, and cross-modal associations in the speech and visual input on the other hand.

The use of the term 'word' in the context of discovery may be a bit misleading, due to the meanings of the term in linguistics. Throughout this chapter 'word' means an entity of which an acoustic realisation is present across utterances as a stretch of speech.

Given a database consisting of 8000 utterances, we showed that our learning model is able to build and update representations of 13 different target words. Experiment 1 shows that these representations are speaker dependent: When the learner is confronted with a new speaker, the model must adapt its internal representation to the characteristics of the speech of the new speaker. A computational model like we are building allows us to look inside the representation space and to investigate the dynamic behaviour of representations during learning.

Experiment 2 showed that the actual performance of the learner depends on two parameters that determine when and how the internal representations are updated. The amount of utterances that is used for each update of internal representations relates to the amount of memory that can be kept active during training. The result in experiment 2 suggests that 10 to 50 observations must be kept in memory for building adequate representations of words. The second result of experiment 2 shows that the amount of data used for bootstrapping the NMF decomposition is not crucial for the eventual performance of the learner. This means that learning can be considered as a truly ongoing process, operating directly from the first stimulus.

The third experiment showed that the representations are changing continuously, and that the representation space can be investigated in detail. A clustering of the columns of W showed how speaker-dependent word representations can be grouped into clusters that correspond almost 1-1 with speaker-independent word representations.
The conceptual consequences of this result are very interesting. In the literature on mental representations of words and the status of phonemes in the prelexical representation (see e.g. McQueen, 2007) there is considerable discussion about the level of abstractness that must be assumed in the word representations. Although based on a simple database and a simple word discovery scheme, the result in experiment 3 suggests how abstraction may follow as a result of competition between crowded collections of representations on a lower level. If needed, speaker-dependent word representations can be clustered such that the common characteristics of these representations combine into one unique word representation.

The current word discovery approach does not use the ordering of the words in an utterance. The utterances 'the ball is red' and 'the red is ball' would be mapped onto the same vector (there are small differences that are not relevant for this discussion). This seems an undesirable property of a word discovery algorithm, especially when the acquisition of syntax is a next step in language acquisition (cf. Saffran and Wilson, 2003). Current research shows that NMF *is* able to recover information about word order by augmenting the output of the map function with additional components related to the relative position of words in the input. A discussion about this approach is outside the scope of this paper.

Since the computational model aims at simulating word discovery as it could happen in human language acquisition, the cognitive plausibility of the model is an important evaluation criterion. The literature on language learning and word acquisition discusses a number of phenomena.
Firstly, the number of words that young infants understand increases over time, with a 'word spurt' between the age 1y and 2y. This word spurt is generally attributed to various factors such as effective reuse of existing representations (but other factors may play a role, see McWhinney, 1998). In the current experiments, a word spurt effect is not yet shown. The way in which internal representations are built, however, paves the way to investigate whether a word spurt effect can be (at least partly) explained by the efficient reuse of already-trained internal representations. If the representations space becomes too crowded, this may be a trigger for the learner to look for a more efficient encoding of the stored information, with a better (more efficient) decoding of new words as a possible result.

In the language acquisition literature, a few more characteristics of language learning are discussed of which the modelling will be a challenge for all models that are ultimately based on statistics. One of these characteristics is that infants reach a stage in which they need just a few examples to learn a new word. Apparently, a reliable representation can be built on the basis of a few tokens only. Our model is in principle able to do that, but to what extent this is dependent on other factors remains to be investigated. Investigations about how a training could be performed on the basis of single tokens (or just a few tokens) will help to understand to what extent the human speech decoding process deviates from a purely Bayesian model.

Another characteristic of first language acquisition is a phenomenon referred to as *fast mapping*. Broadly speaking, fast mapping means that children learn that 'new' (unobserved) words are likely to refer to 'so far unobserved' objects. Apparently the formation of form-referent pairs is a process that might be controlled by some economic rules (in combination with statistically motivated updates of representations). For example, it may imply that an utterance that cannot be understood (fully parsed) given the current representations inspires the learner to postulate a new word-referent pair. However, we want to avoid an ad-hoc approach, in the sense that we want to avoid that the computational model is able to reproduce the effects due to a pre-thought scheme in the implementation. Instead, the fast mapping may result from the use of an underlying rule e.g. based on efficient reuse of representations or on efficient interpretation of the stimulus. The phenomenon of fast mapping will be topic of experiments in the near future.

Our last discussion point relates to the use of visual/semantic tags in the multimodal databases. In the experiments reported in this chapter, tags serve as an abstract representation of the object in the scene that the utterance relates to. The tags are now interpreted by the computational model as they are, without any uncertainty that might obscure its precise interpretation. This might be regarded as undesirable, since it favours the visual information compared to the auditory input (which is subject to variation and uncertainty). Moreover, it is not realistic to assume that the visual system is able to come up with unambiguous and invariant tags.
In the near future the computational model will be extended with a component that allows us to present 'truly' multimodal stimuli, comprising of an audio component and 'visual/semantic' component. The visual/semantic component will then replace the tag that was used in the current databases. For example: the tag 'ball' will be replaced by a vector of binary components, each of them indicated the presence or absence of a certain primitive visual feature (such as red-ness, blue-ness, round-ness).

7 Conclusion

We presented a computational model of word discovery as the first step in language acquisition. The word representations emerge during training without being specified a priori. Word-like entities are discovered without the necessity to first detect sub-word units. The results show that 13 target words can be detected with an accuracy of 95-98 percent by using a database of 8000 utterances spoken by 4 speakers (2000 utterances per speaker).

Future research will enhance the model such that information about word ordering can be obtained. Also the multi-modal information in the stimuli will be enriched to encode visual/semantic information in a cognitively more plausible way.

Acknowledgements

This research was funded by the European Commission under contract FP6-034362 (ACORNS).

8. References

Aslin, R.N., Saffran, J.R., Newport, E.L. (1998). Computation of probability statistics by 8-month-old infants. Psychol Sci 9, pp. 321-324.

Bacchiani, M. (1999). Speech recognition system design based on automatically derived units. PhD Thesis, Boston University (Dept. of Electrical and Computer Engineering) (available on-line).

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

Bosch, L. ten (2006). Speech variation and the use of distance metrics on the articulatory feature space. ITRW Workshop on Speech Recognition and Intrinsic Variation, Toulouse.

Bosch, L. ten, and Cranen, B. (2007). An unsupervised model for word discovery. Proceedings Interspeech 2007, Antwerp, Belgium.

Bosch, L. ten, Van hamme, H., Boves, L. (2008). A computational model of wlanguage acquisition: focus on word discovery. Proceedings Interspeech 2008, Brisbane, Australia.

Bourlard, H., Hermansky, H., Morgan, N. (1996). Towards increasing speech recognition error rates. Speech Communication, Volume 18, Issue 3 (May 1996), 205--231.

Boves, L., ten Bosch, L. and Moore, R. (2007). ACORNS - towards computational modeling of communication and recognition skills , in Proc. IEEE conference on cognitive informatics, pages 349-356, August 2007.

Gaskell, M. G. (2007). Statistical and connectionist models of speech perception and word recognition. In M. G. Gaskell (Ed.), The Oxford Handbook of Psycholinguistics, pp. 55-69, Oxford University Press, Oxford, 2007.

George, D. and Hawkins, J. (2005) A Hierarchical Bayesian Model of Invariant Pattern Recognition in the Visual Cortex. Proceedings of the International Joint Conference on Neural Networks (IJCNN 05).

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. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, Vol. 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.

Hoyer, P. (2004). Non-negative matrix factorisation with sparseness constraints. Journal of Machine Learning Research 5. Pp. 1457-1469.

Johnson, E.K., Jusczyk, P.W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. J Mem Lang 44:548-567.

Johnson, S. (2002). Emergence. New York: Scribner.

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.

Lee, C.-H. (2004). From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition. Proc. ICSLP.

Lippmann, R. (1997). Speech Recognition by Human and Machines. Speech Communication, 22: 1-14.

Luce, P.A and Lyons, E.A. (1998) Specificity of memory representations for spoken words, Mem Cognit.,26(4): 708-715.

Maslow, A. (1954). Motivation and Personality New York: Harper & Row.

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, Vol. 18, 1986, pp. 1-86.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. In M. G. Gaskell (Ed.), The Oxford Handbook of Psycholinguistics, pp. 37-53, Oxford University Press, Oxford, 2007.

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).

Newport, E.L. (2006). Statistical language learning in human infants and adults. Plenary addressed at Interspeech 2006, Pittsburgh, USA (Sept. 2006).

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, Vol. 52, 1994, pp. 189-234.

Norris, D. and McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review 115(2), pp.357-395.

Ostendorf, M. (1999). Moving beyond the beads-on-a-string model of speech. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop. Vol. 1. Keystone, Colorado, USA, pp. 79-83.

Park A., and Glass, J. (2006). Unsupervised word acquisition from speech using pattern discovery. Proceedings ICASSP-2006, Toulouse, France, pp. 409--412.

Pisoni, D. B. and Levi, S. V. (2007). Representations and representational specificity in speech perception and spoken word recognition. In M. G. Gaskell (Ed.), The Oxford Handbook of Psycholinguistics, pp. 3-18, Oxford University Press, Oxford, 2007.

Pitt, M.A., Myung, I. J. and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. Psychological Review, Vol. 109, 2002, pp. 472-491.

Prince C.G. and Hollich, G. J. (2005). Synching infants with models: a perceptual-level model of infant synchrony detection. The Journal of Cognitive Systems Research, 6, pp. 205-228.

Roy, D., and Pentland, A. (2002). Learning words from sights and sounds: A computational model. Cognitive Science, 26 (1), 113-146.

Saffran, J.R., Aslin, R.N., and Newport, E.L. (1996). Statistical learning in 8-month-old infants, Science, 274, December, pp. 1926-28.

Saffran, J.R., Wilson, D.P. (2003). From syllables to syntax: multilevel statistical learning by 12-month-old infants. Infancy 4:273--284.

Scharenborg O., Norris, D., ten Bosch, L., and McQueen, J. M. (2005). How should a speech recognizer work? Cognitive Science, Vol. 29, pp. 867-918.

Shi, R., Oshima-Takane, Y., and Marquis A. (2008). Word-meaning association in early language development. Brain and Cognition. Volume 67, Supplement 1, June 2008, Pages 38-39.

Smith, L.B. and Yu C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition 106(3). Pp 1558-1568.

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.

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.

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.

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

Thiessen, E. D., Hill, E.A., and Saffran J.R. (2005). Infant-Directed Speech Facilitates Word Segmentation. Infancy, 7(1), 53--71

Thiessen, E.D., Saffran, J.R. (2003) When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. Dev Psychol 39:706--716.

Wang, Y. (2003). Cognitive Informatics: A new transdiciplinary research field. Brain and Mind, 4: 115-127.

Wang, Y. (2007). On cognitive informatiocs foundations of knowledge and formal knowledge systems. 6th international conference on cognitive informatics, Lake Tahoe, CA, USA, August 6-8, 2007. pp. 263-272.

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.