Investigating word learning processes in an artificial agent

Michele Gubian*, Christina Bergmann*†, Lou Boves*
*Centre for Language & Speech Technology, Radboud University, Nijmegen, The Netherlands
†International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands

***Abstract*—Researchers in human language processing and acquisition are making an increasing use of computational models. Computer simulations provide a valuable platform to reproduce hypothesised learning mechanisms that are otherwise very difficult, if not impossible, to verify on human subjects. However, computational models come with problems and risks. It is difficult to (automatically) extract essential information about the developing internal representations from a set of simulation runs, and often researchers limit themselves to analysing learning curves based on empirical recognition accuracy through time. The associated risk is to erroneously deem a specific learning behaviour as generalisable to human learners, while it could also be a mere consequence (artifact) of the implementation of the artificial learner or of the input coding scheme.**

In this paper a set of simulation runs taken from the ACORNS project is investigated. First a look 'inside the box' of the learner is provided by employing novel quantitative methods for analysing changing structures in large data sets. Then, the obtained findings are discussed in the perspective of their ecological validity in the field of child language acquisition.

***Index Terms*—5.2 grounding of knowledge and representations, 6.1 language learning, 6.8 statistical learning**

I. Introduction

Language acquisition, arguably a highly complex problem, is approached and solved seemingly effortlessly by young children. During their first year alone, as reviewed by Newman [1], infants learn to pay attention to the distinctive and characteristic features of the ambient language and to ignore features that do not contain information relevant to their native language. It has been shown, as summarised in [1], that infants of 7.5 months can identify 'words' from streams of speech after a short familiarisation phase with the words. However, the way infants of that age spot and store those 'words' cannot be compared one-to-one to how adults process language. Among other things, the identity of the speaker has been found to be part of the 'word' representation. This implies that despite their ability to reliably recognise 'words', 7.5 month olds have not yet discovered all acoustic and linguistic properties that actually characterise a meaningful segment of speech; as a consequence they seemingly store an overabundance of acoustic detail.

Moreover, word learning not only requires segmentation and storage of acoustic information, but also the generation of association of acoustic information to objects, attributes or actions in the real-world context to create meaningful units. When learning such concept-label associations, visual information is usually accompanied by a descriptor embedded in the speech stream, that has to be identified and linked to the accompanying visual scene [2].

Almost by necessity experimental research on the nature and acquisition of language skills in infants usually must rely on overt behaviour, such as head turns or eye movements in response to speech stimuli. Internal processes and representations can only be assessed by inference and based on a number of assumptions that cannot be verified easily. Hence, several theories exist concerning what actually has to be learned, what a child brings to the task of language acquisition and how language learning proceeds (for two opposing views see e.g. [3], [4]).

To test basic assumptions, derive new hypotheses and generate predictions, computational modelling is a viable alternative to experimental studies with infants. As opposed to infants, where it is not possible to directly observe internal representations and processes – neither on neural nor on more abstract levels – computational models allow for insights into their own inner workings. In the ideal case designers have detailed control over the structure of the representations as well as on the computational processes that they build into a model. In this way it should be possible to verify the cognitive plausibility of the model based on its construction in addition to merely analyse the fit between experimental and simulated data by comparing the output of a model with the results of behavioural experiments.

However, the main focus of most computational model lays on simulating a child's performance, sometimes with little consideration of available theoretical and factual knowledge concerning processes underlying children's behaviour. Because the actual algorithms employed to simulate cognitive processes are usually – and necessarily – quite complex, their behaviour may not be entirely predictable (e.g. learning based on non-convex optimisation does not always reach a global optimum), they depend on (too) many parameters whose impact is not always well understood, and as a consequence their output and internal representations might be hard to interpret. Furthermore, computational models often concentrate on a specific process and have to approximate factors that are not at the core of the model.

In infant speech comprehension, the behavioural measurements that lay the basis for most models involve physical responses such as head turns, which are rarely explicitly included in computational models. And even if such observable behaviour would be simulated, the link between 'comprehension'

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

and ensuing action is yet another complication of the model, requiring additional assumptions that are difficult to verify experimentally. Rather, abstract measures such as recognition accuracy serve as a measure of a models' performance and are in turn compared to infant data, leading to very indirect comparisons at best. Finally, although computational models offer the invaluable possibility to inspect their internal mechanisms with virtually no limit on the level of detail, there is always a threshold beyond which zooming in would reveal only facts related to the algorithm implementation, with no possible connection with the human mind and brain [6].

The goal of this work is to delve into the problems brought up above, with a main focus on the comparability of computational models and infants beyond performance measurements. A state-of-the-art computational model, namely one of the operational word-learning models developed in the ACORNS project (www.acorns-project.org) which is briefly described in Sec. II, will be studied in depth using a set of simulations of learning word-concept association in infants. Based on the results, we assess the cognitive plausibility of the model's input-output relations and of the dynamics of its internal representations. To this end, we devise an array of measurements that go beyond the analysis of learning curves in Sec. III to allow investigation of the internal representations and their effect on the input-output relations. Those measurements are often indirect and non-trivial, since the internal functioning of the model is not easily interpretable and inherently model-specific. Both issues underline the need to take interpretability in a wider sense than just performance measures into account when designing computational models. When relating our findings on the computational model to existing knowledge about word-learning in infants in Sec. IV, we focus on the studies reviewed in [1] and shortly introduced above.

Overall, our analysis of an existing model sheds light on possible similarities and differences when comparing computational models to infant data. Taking one step back from the specific model we investigated, our data suggest that there is an urgent need to focus on processes and internal representations next to performance when computationally modelling infant word-learning. This is shown by the need to devise specific tools for analysing the model we selected, as well as by the difficulty to distinguish model-specific properties we found from phenomena that emerged from the modelled process and can be generalised to infants.

II. The ACORNS Model

The ACORNS project aimed at modelling language acquisition during cross-situational, multi-modal learning, that is aided by a child's general ability to detect recurrent patterns. The learning process is simulated by a computational model made publicly available by the ACORNS project, where input is presented to a simulated Learning Agent (LA) by a simulated Caregiving Agent (CA) in a multi-modal manner. More precisely, the input consists of an auditory part, a spoken utterance (e.g. 'Look at the ball'), accompanied by a conceptual, pseudo-visual (in the line of [2]) representation of a referent, or *keyword*, that occurred in the sentence (the object 'ball'). No lexical, phonetic or phonological information is provided to the LA, nor is information on the number of different items in the input given beforehand (meaning that the model does not know a priori how many internal representations must be learned).

A. The Computational Model

To fully motivate the analysis tools we develop in the subsequent sections, we provide some background information about the ACORNS computational model we have used, limiting the detail to the minimum which is necessary to understand the technical analysis that follows. For more information the reader is referred to the ACORNS literature and the companion website at www.acorns-project.org.

The learning algorithm used in this particular ACORNS model is Non-negative Matrix Factorisation (NMF) [7]. Inputs are coded as columns v of predefined length n and organised into an $n \times m$ matrix V. The acoustic part of the input V_a holds the first n_a rows of V, while the lower part V_c contains the associated conceptual information associated with each acoustic representation. Learning consists of finding a compact decomposition of V:

$$V \approx W \cdot H \tag{1}$$

where W is of size $n \times r$ and H is $r \times m$, with r being chosen such that $(m+n)r < mn$, i.e. information is (substantially) compressed. Note that due to the product form (1) the organisation of the columns of W is the same as those of V, i.e. they consist of a concatenation of the acoustic part W_a and the conceptual part W_c. The optimal decomposition is chosen by minimising the Kullback-Leibler (KL) divergence between $W \cdot H$ and V. The particular version of NMF used here, which updates the content of W after each input utterance (i.e. each successive column in V), has previously been described in [8]. This update procedure simulates incremental causal learning: The LA can update its internal representations (memory) after each observation, while being unable to use information that will only become available in the future.

To assess the input-output performance of a model during and after training, only the acoustic part v_a of a new utterance containing a previously learned keyword is given without providing the conceptual part v_c. The latter has to be reconstructed by approximating v_a by $W_a \cdot \hat{h}$, where this time only $\hat{h}$ is estimated (again by minimisation of KL-divergence). The same vector $\hat{h}$ is then used to reconstruct the conceptual part by $W_c \cdot \hat{h}$. This reconstruction is then compared against the original information in order to establish whether the correct keyword was recognised.

During and after learning, any time a stimulus v is presented to LA it is internally represented by the vector $\hat{h}$, which contains the (non-negative) proportions of columns of W necessary to optimally reconstruct v. In this respect, $\hat{h}$ can be seen as the analogue of a short term memory, i.e. the pattern of internal representation activations that is produced as a stimulus is received. On the other hand, W permanently

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

stores conceptual-acoustic patterns that come from and have the same structure as the training columns in V. This allows interpreting W as long-term memory (as suggested e.g. in [9]).

In all ACORNS computational models conceptual, pseudo-visual information is symbolic. In our simulations, a pseudo-visual vector v_c has length N_{key} and encodes a specific referent by placing a *one* in the assigned keyword position and *zeros* elsewhere. The content of the sub-matrix W_c can be interpreted in the same way, up to a multiplicative factor.

The encoded acoustic information, on the other hand, comes from real speech. Each acoustic vector v_a has length $n_a = 110,002$ and it is based on a Vector Quantisation coding of the MFCC vectors derived from an input utterance. This high dimensionality is a consequence of the coding scheme that captures co-occurrences of acoustic events at specified time lags [12]. Note that it is not possible to resynthesise the original speech signal from a vector v_a.

B. The Simulations

The simulations described here, which were closely matched to previous ACORNS experiments (as described e.g. in [9], [10]), form the basis of our investigation into the model.

Two word acquisition simulations were conducted, whose motivations and outcomes were previously discussed in [10]. In short, the experiments were designed to test the hypothesis that the LA creates more general internal representations when learning from several speakers than when learning from a single speaker. Both experiments used the same training set selected from the English part of the ACORNS database, namely a collection of $m = 480$ sentences, each one containing one out of $N_{key} = 10$ keywords. The number of columns of W was $r = 70$, which allows room for possible internal organisation beyond a one-to-one mapping with the 10 keywords.

Sentences are short and have a simple structure, which is in accordance with findings concerning child-directed speech [11]. Four speakers, two female and two male, assume the role of caregiver. To investigate whether speaker specific representations will emerge if the learner interacts with each speaker in sequence, we ran two simulations. In the *speaker-mixed* simulation the occurrence of each speaker and keyword was randomised, yet balanced for repetition. In the *speaker-blocked* simulation, the learner was first taught by the first speaker, then by the second one, and so on, while the sentence order was randomised within each block.

A held out test set containing all keywords spoken three times by each speaker was used to measure recognition accuracy. The same test set was used repeatedly, after each set of 10 training utterances. During testing, the incremental learning was switched off. Thus, the LA does not remember anything about the test set. This means that the same test set can be used repeatedly, making it possible to create learning curves, which show the percentage of correctly recognised stimuli as a function of the number of training utterances.

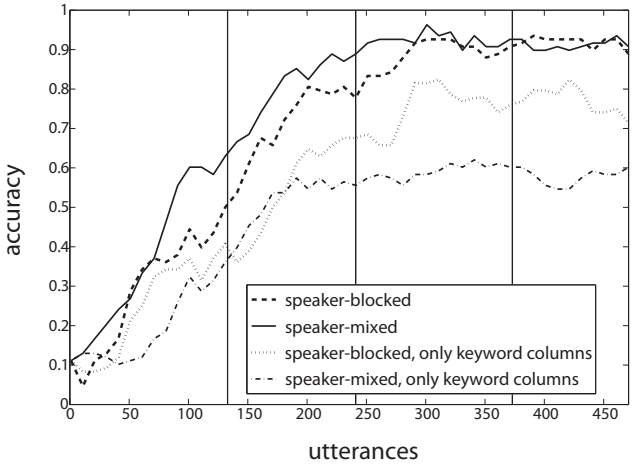

Fig. 1: Accuracy in both the *speaker-mixed* and the *speaker-blocked* conditions with either complete W_a (solid line for *speaker-mixed* and dashed line for *speaker-blocked*) or a limited set of only keyword-encoding columns (dotted-dashed line for *speaker-mixed* and dotted line for *speaker-blocked*). The horizontal lines indicate the onset of a new speaker in the *speaker-blocked* condition.

III. Analysis of the Simulations

In this section we analyse the simulations described in Sec. II-B in depth. In doing so, we will not limit ourselves to inspect learning curves. Instead, we try to look 'inside the box' of the learning algorithm in order to get the necessary insight that will be related to experimental findings on infants in the next section. To this end, a number of additional measurements beyond accuracy will have to be chosen, as the inner workings of a model are not completely transparent.

A. Learning Curves

The learning curves for the simulations described in Sec. II-B can be inspected in Fig. 1 (solid and dashed line respectively). It can be seen that learning proceeds gradually, and that the two conditions (*speaker-mixed* and *speaker-blocked*) perform on a similar level of accuracy after about half the training set has been observed. From the similarity between the learning curves for the two conditions it can be inferred that the system is able to 'understand' all four speakers, even if it has not yet been trained with speakers 2, 3 and 4 in the *speaker-blocked* condition.

B. Learning in the Conceptual Memory W_c

In Fig. 2 the content of the pseudo-visual memory W_c in the *speaker-blocked* simulation at the end of the training phase is shown. At most one keyword is encoded in a column and most keywords are represented by a unique column. This result suggests that there is no tendency to produce episodic representations, since, for example, there is no evidence of speaker dependent representations.

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

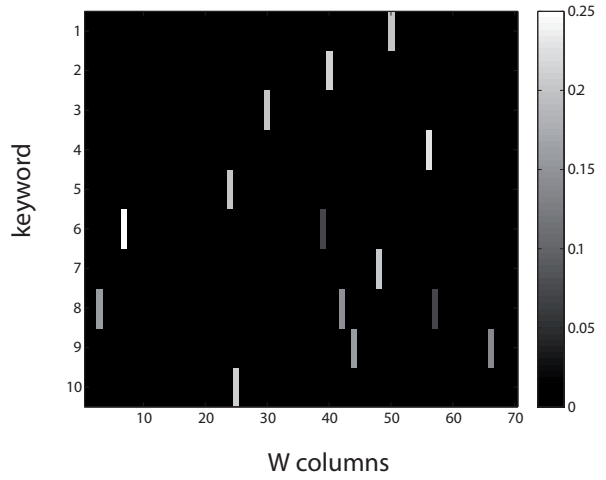

Fig. 2: Internal representation of the conceptual information stored in W_c for the *speaker-blocked* condition at the end of training.

By playing the sequence of all $W_c(t)$ snapshots, $t = 1, \ldots, m$ to study the emergence of associations between auditory and conceptual representations, clear activations in specific columns appear after all keywords have been presented only a few times. Columns associated with keywords tend to sharpen their peaks and no oscillations between columns or instabilities are visible. Similar results were found in the *speaker-mixed* condition.

C. Learning in the Auditory Memory W_a

Inspired by the findings described above, we wanted to investigate whether the organisation of the pseudo-visual memory W_c is replicated in the auditory memory W_a, i.e., whether there is a small subset of columns that encodes the keywords, while the rest of the space is (apparently) not used to represent associations between speech and meaning. Just like for W_c, the W_a columns have the same form as the audio input vectors v_a. Therefore, we expect to contain keyword-encoding columns in the visual part W_c to contain a corresponding, keyword-related acoustic association in W_a. However, since it is not possible to resynthesise the original speech signal from a vector v_a we need to develop a more indirect approach for investigating the structure of the internal representations of the acoustic part of the 'speech-meaning' associations.

If some columns in W_a encode keywords, we expect them to exhibit sharp peaks denoting the presence of the sound patterns characteristic for those words, with different words creating peaks in different positions of a vector. Columns with no specific sound-keyword association are likely to have a more uniform noise-like appearance. In order to investigate this hypothesis, we adopted the following measure of dissimilarity between two acoustic columns p and q:

$$d(p,q) = \sum_{i=1}^{n_a} \overline{p}_i \log \frac{\overline{p}_i}{\overline{q}_i} + \overline{q}_i \log \frac{\overline{q}_i}{\overline{p}_i} \qquad (2)$$

i.e. a symmetric version of the KL-divergence between vectors, where $\overline{x_i} = \frac{x_i}{\sum_{i=1}^{n_a} x_i}$. If p and q exhibit peaks in coinciding positions, $d(p,q)$ tends to be less than one; peaks in different positions lead to $d(p,q) > 1$; if p and q contain uniformly distributed and uncorrelated noise, then it can be shown that $d(p,q)$ tends to one as n_a tends to infinity.

Using (2) we built a dissimilarity matrix D over all the $\binom{r}{2}$ column pairs of W_a. We used a hierarchical clustering algorithm based on the dissimilarities in D to infer the underlying structure of W_a. We expected to uncover the presence of ten singleton or two-member-clusters containing keyword-encoding columns and a big cluster containing the remaining columns of W_a.

To leave it to the data to determine the number of clusters K with the best fit, we calculated the average *silhouette* value s for $1 \leq K \leq r = 70$ (the maximum value of K corresponds to the situation that each column is a cluster in its own right) [13]. The silhouette value of a cluster element is an empirical index in $[-1, 1]$ denoting how well that element is contained in its own cluster. The average s over all 70 elements provides a global 'fitness' value for the clustering. We computed s for each possible value of K as well as for each learning step $t = 1, \ldots, m$. Fig. 3 shows a grey scale map $s(t, K)$ for the *speaker-mixed* condition. Values of s around 0.5 and higher are considered to be trustworthy and are found from $K = 2$ (by definition $s = 0$ for $K = 1$) up to around $N_{key} + 1$ from very early in training process onwards. We also verified manually (i.e. by imposing $K = N_{key} + 1$ at several points in time) that N_{key} clusters indeed contained the same columns that exhibit peaks in the W_c part, in addition to a big and diffuse cluster collecting the remaining columns (Fig. 2).

The lack of a clear preference for $K = N_{key} + 1$ in comparison to lower K may be attributed to the nature of the dissimilarity (2), which does not satisfy the triangular inequality. As a consequence, the overall s does not change substantially if a singleton cluster representing a keyword is merged with the diffuse cluster formed by the non-keyword columns. A similar pattern for s was found for the *speaker-blocked* case, which confirms the absence of systematic speaker-dependent internal representations.

D. Evolution of Keyword Representations

The cluster analysis gave an initial impression of the content of W_a and the effect of training in terms of the number of elementary units related to keywords. Because of the relatively stable accuracy scores and number of clusters, a reasonable expectation would be that the keyword-encoding columns of W reach a stable state very early and get updated mostly upon presentations of utterances containing their keyword. No hypothesis could be elaborated on the behaviour of the other columns.

To inspect the changes that each column undergoes throughout learning, eq. (2) was applied to each pair of points in time (t_1, t_2) for each column separately, i.e., computing how column p at time t_2 differs from itself at previous time t_1. Representative results for a word-encoding column are shown

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

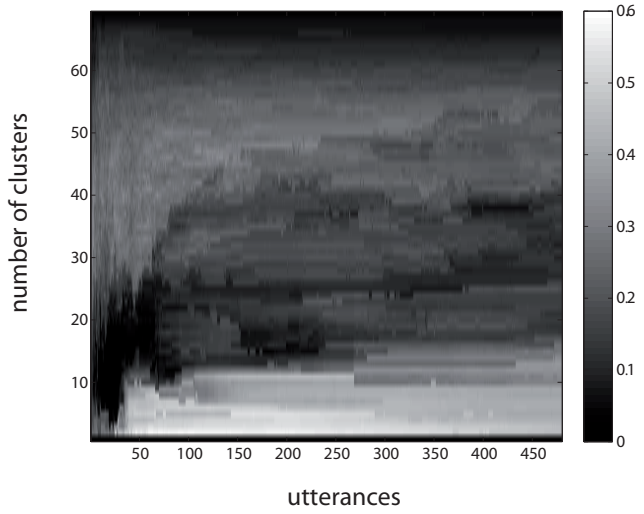

Fig. 3: Average silhouette value s for each learning step t (x-axis) and each number of clusters K (y-axis) for the *speaker-mixed* condition.

in the grey scale maps $d(p(t_1), p(t_2))$ in Fig. 4a and 4b for the *speaker-mixed* and the *speaker-blocked* case, respectively. Visual inspection of the two figures reveals two phenomena. First, columns do not go back to previous configurations; rather, they continue to evolve (all horizontal or vertical cuts in $d(p(t_1), p(t_2))$ are V-shaped, with the minimum at $t_1 = t_2$ and an increase of d when moving away from the minimum). Second, the pixel-like appearance that can be seen in Fig. 4a coincides with the presentations of the corresponding keyword, meaning that columns react only to their own keyword. Moreover, a macro-blocked structure is visible in Fig. 4b, which coincides with the speaker changes during training. It seems that an incoming new speaker induces a strong reaction in the system, which leads to adjustment of each existing internal representation by the learning engine without the need to create a new one.

Two $d(p(t_1), p(t_2))$ maps of non-keyword columns are shown in Fig. 4c and 4d for *speaker-mixed* and *speaker-blocked* case, respectively. While the continued evolution is found here as well, no particular structure is visible in those maps, with the exception of a clear reaction to the incoming third speaker in Fig. 4d. Therefore, we are still not able to formulate hypotheses about the function of the non-keyword columns during and after learning.

The first result of this investigation is that there is seemingly no stable state both for keyword or non-keyword columns within the observed training time, because columns do not return to the same configuration. Second, the keyword columns update upon encountering examples of the encoded keyword, which leads to step-wise shifts through the space as opposed to the smooth transitions visible in the non-keyword columns. Furthermore, the speaker-change leads to grater changes than presentations of the same keyword by the same speaker in different carrier sentences. Hence, some information about the speaker must have been part of the acoustic representation.

E. Reconstructing Auditory Input – The Role of Non-Keyword Columns

Are non-keyword columns used at all in reconstructing utterances? We tried to answer this question by applying the accuracy measurement described in section Sec. II-A using only keyword columns of W_a for recognising test stimuli. The results are depicted in Fig. 1 as the dash-dot and dotted lines, together with the original accuracy scores. The results show that recognition accuracy suffers substantially when a given sentence has to be approximated by only the keyword columns in W_a. This holds for both the *speaker-mixed* and the *speaker-blocked* condition. This finding rules out the hypothesis that non-keyword columns are simply not used or not useful. Therefore, we must assume that they encode acoustic elements related to the carrier sentences, possibly associated to frequent words or word groups, or perhaps associated to characteristic voice qualities of the speakers.

We attempted to discover the function of the non-keyword column by creating a linear regression model whose inputs are binary (dummy) predictors describing an input utterance by the presence or absence of keywords, frequent words or sentence fragments, and gender and identity of the speaker. The output is the value of the coefficient in $\hat{h}$ corresponding to a specific column in W_a when an utterance is reconstructed by the learning algorithm.

Keyword column outputs were very well explained just by their keyword predictor ($R^2 \approx 0.8$). Non-keyword column models were hard (if at all) to interpret, and the explained variance was seldom above $R^2 \approx 0.2$. Manual inspection of the linear models only brought out effects that were due to idiosyncrasies in the training set, e.g. non-keyword columns showing moderate effects of a word and one particular speaker, when the word was pronounced by this one speaker alone in the training set. We believe that the failure to find interpretable structure in the non-keyword columns is not due to the specific choice of the inspection tool (classic linear model) but that it is related to the choice of the predictors. The coding scheme implemented in W_a is very close to the acoustic signal, while our predictors are at a high level of (linguistic) abstraction. The limited size of the training set probably does not allow for high level representations like words or phones to emerge in an unsupervised setting. The exception of keywords is explained by the fact that they are learned with supervision.

F. Discussion

To summarise the main findings of the previous section in the order they were presented above, we first can note that the learner is able to correctly associate sounds to keywords early in the process and with good generalisation capacities, as shown in the learning curves in Fig. 1. Investigating the internal organisation of the learner's memory we could reveal the presence of memory locations (columns) dedicated to the association of one single keyword to an acoustic pattern. A cluster analysis showed that those acoustic patterns associated to a particular keyword differ very much from those associated

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

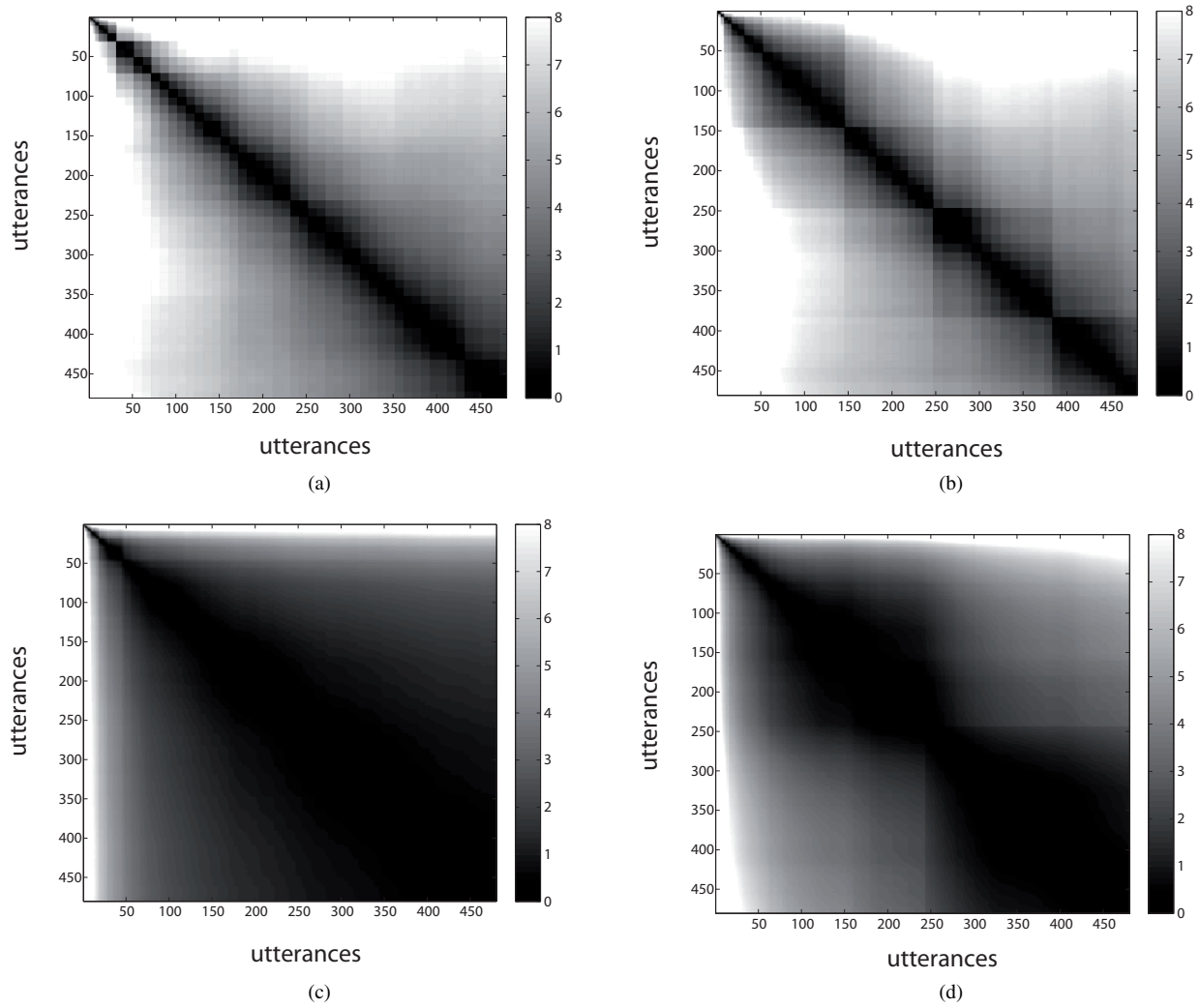

Fig. 4: Dissimilarity measure (2) between an acoustic column of W_a at time t_1 and the same column at t_2. A keyword column in (a) *speaker-mixed* and (b) *speaker-blocked* condition. A non-keyword column in (c) *speaker-mixed* and (d) *speaker-blocked* condition.

with other keywords and it plays a dominant role in the recognition (reconstruction) of input containing that keyword.

The concept-sound association appears in the memory after only a few presentations of the relevant paired stimuli. After the emergence of these columns, no major memory reorganisation was encountered. Still, the system continued to adapt its representations to the incoming new learning stimuli. Even well after recognition accuracy reaches ceiling, a column dedicated to a specific keyword keeps being modified by incoming input containing the same keyword (Fig. 4). No evidence of emerging speaker-dependent representations was found but the adaptation that a keyword-column undergoes when a new speaker is introduced was stronger than other updates in the same simulation or general changes in the *speaker-mixed* condition.

Our attempts to understand the role of the memory locations not bounded to keywords did not bring any clear interpretation. They are useful in the recognition of audio but they don't seem to code anything that we can interpret. Hence, we will exclude them from the subsequent discussion and leave this topic open for further investigation.

IV. Relating the Simulation Results to Infant Data

The findings of the technical analysis of the simulations above have to be related to findings from experiments on language acquisition in infants. First, it should be noted that ACORNS only aims at modelling a simplified and highly constrained word-learning task, which constitutes a subset of the tasks a child is confronted in his or her first year. Moreover, the amount of input given to the model is comparable to the number of sentences a child hears within a few days of his or her life in infant-directed speech, as found by [11].

978-1-4244-6902-4/10/$26.00 ©2010 IEEE

Having this in mind, the main finding from the technical analysis is the fast and stable one-to-one binding of acoustic internal representations to the pseudo-visual counterpart when encoding a keyword. On one hand, this fact resonates well with experimental evidence in child language acquisition in that this fast recognition of familiar keywords can be found in infants too [1]. However, testing the fast formation of such internal representations in the lab can be disturbed by a number of experimental factors, and is consequently not as robust as the present findings might suggest [14]. Additionally, unlike in our simulations, children rapidly forget such word-object mappings when they were only encountered a few times or in an experimental setting with high cognitive load. This property of the child's memory plays a crucial role in both experimental findings and during day-to-day language learning and cannot easily be captured by the present model. Forgetting can be implemented in the present model. However, such an implementation is all but trivial, if only because several different technical options are available, each implying a different hypothesis about how 'forgetting' works in the infant brain.

If we then look back to the ACORNS model mechanics, we can see that even though the one-to-one associations were emergent and not imposed, the pseudo-visual coding is so powerful due to its orthogonality that the system is strongly biased to this kind of organisation, and other more sophisticated ones are unlikely to appear. Any remedy for this seems to depend on the choice of conceptual or pseudo-visual coding. As there was for example no speaker-dependent encoding given to the system, no specific memory locations for each speaker could emerge. Still, a strong reaction to changing the speaker was observable, which is indicative for a detection of inherent differences. Again, this can be seen as an artifact of the coding, with the speaker change reactions are the only possible emergent behaviour that is allowed. Hence, we can assume that indeed also speaker-dependent behaviour was found, but in a way that was be masked by the way this particular model encodes accompanying information in non-acoustic modalities.

V. Conclusions

Our results demonstrate that it is difficult to examine a specific computational model in detail, as well as how difficult it can be to relate the results of computer simulations to what is being modelled, namely infant word-learning. Specific additional tools to closely examine the inner workings of the ACORNS model had to be developed, as they were not part of either the model or the ACORNS project.

Furthermore, it was difficult to tease apart effects which hold in general from effects that derive from specific choices in the technical implementation of the model. This was partly due to properties of the learning algorithm, which is based on matrix decomposition and might lead to observations such as those mentioned in Sec. III-E. A further source of possible idiosyncratic effects was in the encoding of the input, which consisted of continuous audio input and an abstract, symbolic labelling of keyword-related information. We hypothesised in the section IV that changes in the conceptual coding scheme will lead to different observations within the model's memory structures. Additionally, the strong binding of acoustic and conceptual information found from very early on in the training in Sec. III seems to mainly stem from the orthogonality of the encoding. Hence, there is a direct effect of the form of representation of the non-acoustic input. However, the full extent of that impact cannot be fully understood from the limited amount of experiments conducted within this paper and would require further investigation.

Overall, we can emphasise the need for a detailed inspection of a computational model, which includes an examination of its inner workings. To this end, it would be necessary to either chose a transparent model or provide tools to enable this inspection. With such tools, the assessment of a model, both in terms of functionality and with respect to its ecological validity, would be simplified.

References

[1] R. S. Newman, "The level of detail in infants' word learning," *Current Directions in Psychological Science*, pp. 229–232, 2008.

[2] L. Smith and C. Yu, "Infants rapidly learn word-referent mapping via cross-situational statistics," *Cognition*, vol. 106, pp. 333–338, 2008.

[3] M. Tomasello, *Constructing a language – a usage-based theory of language acquisition*. Harvard University Press, 2003.

[4] N. Chomsky, *New Horizons in the Study of Language and Mind*. Cambridge University Press, 2000.

[5] G. Aimetti, L. ten Bosch, and R. K. Moore, "The emergence of words: Modelling early language acquisition with a dynamic systems perspective," in *Proceedings of the Ninth International Conference on Epigenetic Robotics*, Venice, Italy, 2009, pp. 17 – 24.

[6] O. Scharenborg and L. Boves, "Computational modelling of spoken-word recognition processes: Design choices and evaluation," *Pragmatics & Cognition*, vol. 18, pp. 136–164, 2010.

[7] D. Lee and S. Seung, "Learning the parts of object by non-negative matrix factorization," *Nature*, vol. 40, pp. 788–791, 1999.

[8] V. Stouten, K. Demuynck, and H. Van hamme, "Automatically learning the units of speech by non-negative matrix factorisation," in *Proceedings Interspeech 2007*, Antwerp, Belgium, 2007.

[9] L. ten Bosch, H. Van hamme, L. Boves, and R. Moore, "A computational model of language acquisition: the emergence of words," *Fundamenta Informaticae*, pp. 229–249, 2009.

[10] J. Driesen, L. ten Bosch, and H. Van hamme, "On a computational model for language acquisition: modeling cross-speaker generalisation," in *Proc. Text, Speech and Dialogue, 12th Intern. Conference*, 2009.

[11] J. van de Weijer, "Language input for word discovery," Ph.D. dissertation, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, 1998.

[12] H. Van hamme, "HAC-models: a Novel Approach to Continuous Speech Recognition," in *Proceedings Interspeech 2008*, Brisbane, Australia, 2008.

[13] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley, 1990.

[14] J. F. Werker, L. B. Cohen, V. L. Lloyd, M. Casasola, and C. Stager, "Acquisition of word-object associations by 14-month-old infants," *Developmental Psychology*, vol. 34, pp. 1289–1309, 1998.

[15] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," *Cognitive Psychology*, vol. 47, pp. 204–238, 2003.

978-1-4244-6902-4/10/$26.00 ©2010 IEEE