The emergence of words: Modelling early language acquisition with a dynamic systems perspective

Guillaume Aimetti* **Louis ten Bosch**** **Roger K. Moore***

*Speech & Hearing Group
University of Sheffield
Sheffield, UK
{g.aimetti|r.k.moore}@dcs.shef.ac.uk

**Department of Linguistics
Radboud University
Nijmegan, NL
l.tenbosch@let.ru.nl

Abstract

This paper introduces a computational model of early language acquisition that is able to build word-like units from cross-modal stimuli (acoustic and pseudo-visual). The architecture, data processing and internal representations of the model strives for ecological plausibility, and is therefore inspired by current cognitive views of preverbal infant language learning behaviour. In this paper, we attempt to visualise the emergence and development of the models internal representations as an epigenetic landscape, which is a popular method for depicting the evolution of behaviour through the dynamic systems theory. We show that our computational model, through a general statistical learning mechanism, displays similar properties to the dynamic systems theory and supports the empiricist view of human development.

1. Introduction

An increasingly popular view, of developmental researchers, is that the brain is a complex dynamic system and behaviour is emergent through self-organization, known as the dynamic systems theory (DST) (Kelso, 1995, Muchisky et al., 1996, Newell et al., 2003, Smith and Thelen, 2003, Evans, 2007). This perspective takes an empiricist view of development, stating that the acquisition of behaviour is based on a general statistical learning mechanism which is dependent upon experience and initial control parameters. The set of behavioural states of the brain defines a landscape: "Development, then, can be envisioned as a changing landscape of preferred, but not obligatory, behavioural states with varying degrees of stability" (Thelen and Smith, 1995). This view of development, as a constantly evolving landscape, challenges the nativist view that infants are 'hard-wired' with skills that are at their disposal from birth or appear at discrete, arbitrary time-steps. As an example, nativists suggest that young language learners are born with an innate language acquisition device, a universal grammar, which allows them to derive the structure of their native language during a critical period of infancy (Chomsky, 1975, Pinker, 1994).

In the DST framework, attractor states emerge and strengthen as a result of the repeating patterns of the co-operative actions of the systems components. Learning can thus be seen as a shift or bifurcation into a new attractor state by the destabilisation of older stable states (Thelen and Smith, 1995). Behaviour is classed into more or less stable attractor states and changes between these states have a non-linear relationship with environmental input. The behaviour of the system becomes more complex with age, with the formation of multiple attractor states. The wider areas encompass certain categories of actions such as walking, jogging and sprinting.

The timing of developmental changes is controlled by variation in the control parameters, body or environmental changes, rather than some kind of internal clock. Thelen strengthened this theory, overturning the previously held belief that developmental changes were due to cortical inhibition, by proving that the stepping reflex in newborns disappears due to an increase in non-muscular body mass and then reappears when the legs are, once again, strong enough (Thelen and Fisher, 1982). This sparked further research into the application of DST to other motor skills, such as the development of motor skills required to reach for an object (Savelsbergh and Van der Kamp, 1993). DST can thus be used to predict the behaviour of a system with varying control parameters. Thelen argues that the view of development as an evolving landscape is not supposed to prescribe behaviour, but represent a probability of behaviour of a system with varying control parameters.

The epigenetic landscape is currently a popular method for visualising behavioural evolution within developmental science, and was originally drawn in 1957 to display the developmental stability of phenotype over time (Waddington, 1957). Figure 1 is

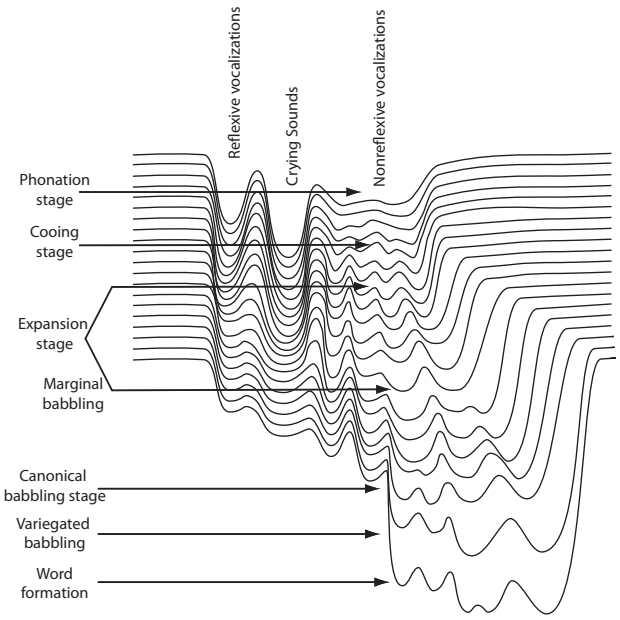


Figure 1: *Diagram of the evolving speech production attractor landscape as illustrated by (Muchisky et al., 1996).*

a diagram of the attractor landscape for the acquisition of speech production skills of an infant as envisioned by current developmental theorists (Muchisky et al., 1996). The three dimensions represent a) time, b) emergent behaviour, and c) the relative stability of the system at any point in time. Each attractor well is a state of behaviour. The deeper the well of an attractor, the more stable the system is when in that state.

It is becoming commonplace to analyse connectionist models, particularly recurrent neural networks, as dynamic systems. We use DST to analyse our computational model in an attempt to gain a deeper understanding of the dynamically evolving internal representations.

The paper is organised as follows. The next section introduces the main components of our computational model, followed by a keyword detection experiment and results. The penultimate section analyses the internal representations through the DST theory. The final section concludes the work and discusses future work being carried out.

2. The computational model

This section describes the Acoustic DP-ngram algorithm (Aimetti, 2009), which is one of three alternative implementations of a comprehensive model of early language acquisition under development in the FP6 FET project ACORNS[1]. The other two methods are Non-negative Matrix Factorisation (NMF) (Stouten et al., 2007) and Concept Matrices (CM) (Räsänen et al., 2009). CM is the most symbolic approach, detecting recurrent patterns of discrete framed-based codebook labels. DP-ngrams is the most episodic, finding repeating patterns from the raw acoustic signal. NMF sits between the two. Another difference is that CM and DP-ngrams take into account the dynamics of the speech signal over time, whereas NMF does not. Instead, NMF processes the whole utterance to form a representation in memory and at a later stage decomposes it to discover structure in the signal.

Figure 2 displays the interactive framework between the carergiver (carer) and learning agent (LA), along with LA's learning processes within a cognitively motivated memory architecture (Jones et al., 2006).

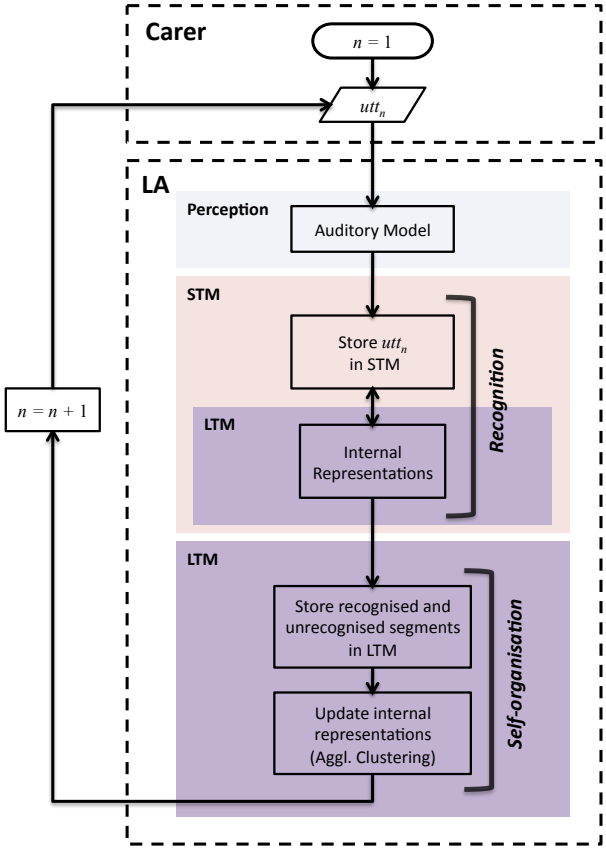


Figure 2: *Flowchart of the carer-learner interactive framework and learning process within a cognitively motivated memory architecture (Jones et al., 2006).*

LA is incrementally presented cross-modal utterances ($utt_{1:\infty}$) by the carer, which contain the raw acoustic signal and a pseudo-visual representation of a keyword (represented as a canonical binary feature indicating it's presence within the utterance). Each utterance contains one of ten keywords (bath, telephone, mummy, daddy, car, bottle, nappy, shoe, book and Angus) and has been constructed using a

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

simple syntax, such as 'Have you seen the W?', where 'W' is a keyword. LA carries out *recognition* using its internal representations. By this procedure, each utterance (utt_n) is segmented into recognised and unrecognised acoustic segments which are appended to long-term memory (LTM). The segment list in LTM is denoted by $X = \{x_1, \ldots, x_m\}$.

Internal representations of keywords and non-keywords emerge through *self-organisation* as a result of clustering the elements of X, on the basis of acoustic similarity, and accumulating their associated pseudo-visual features. The learning processes are discussed in more detail in the following sections.

2.1 Automatic acoustic segmentation

Automatic acoustic segmentation is carried out using the Acoustic DP-ngram algorithm (Aimetti, 2009). This algorithm is a modification of two previous DP-ngram implementations, the first of which was used to find sub-repetitions within a gene sequence (Sankoff and Kruskal, 1983), and the second was used to find sub-repetitions of the output of a phonetic transcription (Nowell and Moore, 1995). The two previous implementations are limited to sequences of discrete symbols, whereas the new implementation can handle multi-dimensional feature vectors. When carrying out experiments directly on the raw acoustic signal we parameterise it to a series of 39-dimensional mel-frequency cepstral Coefficients (MFCC's), which reflect the frequency sensitivity of the human auditory system.

This Acoustic DP-ngram method uses a popular dynamic programming technique, dynamic time warping (DTW), in order to accommodate temporal distortion present in the acoustic speech signal (similar approaches include (ten Bosch and Cranen, 2007, Park and Glass, 2008)). Through an accumulative scoring mechanism, this method is able to detect similar portions of speech that commonly re-occur within utterances (such as phones, words and sentences) whilst being robust against noise, speech rate and pronunciation variation. The discovered sub-sequence portions are termed *local alignments*. An additional property of the accumulative quality score is that longer, more meaningful local alignments produce a higher final quality score, thus allowing the system to list them in order of importance. The three steps of the segmentation process are outlined below.

Step 1: The carer presents LA with the n^{th} utterance (utt_n), which is stored in short-term memory (STM) as a set of MFCC feature vectors (A). LA then carries out template based recognition by comparing this input representation with each internal representation (B). Both A and B are represented as sequences of feature vectors. By applying the Euclidean Squared Distance between each pair of feature vectors (v_A, v_B) we obtain a distance matrix $D = (d(v_A, v_B)v_a, v_b)$.

Step 2: D is then used to calculate the accumulative quality scores for successive frame steps within A and B using the recurrence defined by (1) to give the global quality score matrix Q. Higher local quality scores $q_{i,j}$ are obtained by accumulating successive local-matches, thus the score for a local-match must be positive, and scores for non-matches (insertions and deletions) must be negative to penalise temporal distortion (2).

$$q_{i,j} = max \begin{cases} q_{i-1,j-1} + (s(a_i, b_j) \cdot d(v_i, v_j)), \\ q_{i,j-1} + (s(\phi, b_j) \cdot |d(v_i, v_{-j}) - 1| \cdot q_{i,j-1}), \\ q_{i-1,j} + (s(a_i, \phi) \cdot |d(v_{-i}, v_j) - 1| \cdot q_{i-1,j}), \\ 0 \end{cases} \tag{1}$$

where,

$$\begin{array}{lll} s(a_i, b_j) & = +1 & \text{(local-match score)} \\ s(\phi, b_j) & = -1 & \text{(insertion score)} \\ s(a_i, \phi) & = -1 & \text{(deletion score)} \\ q_{i,j} & & \text{(local quality score)} \end{array} \tag{2}$$

Backtracking pointers p are maintained at each step of the recursion (3).

$$p_{i,j} = \begin{cases} (i-1, j-1), & \text{(local-match)} \\ (i, j-1), & \text{(insertion)} \\ (i-1, j), & \text{(deletion)} \\ (0, 0) & \text{(initial pointer)} \end{cases} \tag{3}$$

Step 3: Finally, the optimal local alignment is discovered within Q by backtracking from the highest quality score $max(q_{i,j})$ until $q_{i,j}$ equals 0. Multiple local alignments are discovered by repeating this process while $max(q_{i,j})$ is greater than the quality threshold (q_{thresh}).

2.2 The emergence of meaning

The incoming utterance is presented to the system in two modalities in parallel, acoustic and pseudo-visual. The pseudo-visual stream contains keyword information as a canonical representation, each keyword is assigned a binary value indicating whether it's present or not present within the current utterance. It is important to note that there is *no* lexical or phonetic information attached to the pseudo-visual feature and no a priori knowledge is assumed. As the incoming utterance is segmented into recognised and unrecognised portions, LA is also associating co-occurring pseudo-visual features to them.

The next section shows an example of the associative learning process for the first two utterances; for the sake of clarity we are using orthographic and not acoustic data for these examples:

1. Begin life

utt_1	
Acoustic	Visual
'the_bottle_is_on_the_seat'	0 0 0 1 0 0

LA does not recognise any of the utterance as there are no internal representations yet, so utt_1 is stored in LTM as a token in cluster C_1.

LTM		
C	Segments	Visual
1	the_bottle_is_on_the_seat	0 0 0 1 0 0

2. Next utterance

utt_1	
Acoustic	Visual
'have_you_seen_the_bottle'	0 0 0 1 0 0

LA compares utt_2 with the internal representation C_1 and recognises the acoustic segment 'the_bottle' and associates it with the co-occuring visual feature. The recognised segment is stored as a token in C_2.

LTM		
C	Segments	Visual
1	the_bottle_is_on_the_seat	0 0 0 1 0 0
2	the_bottle	0 0 0 1 0 0
3	have_you_seen_	0 0 0 0 0 0

The unrecognised portion of utt_2 is stored in C_3 with no associated visual features as it has already been recognised and associated with C_2.

The associative learning mechanism implemented within this algorithm has been cognitively motivated by current developmental theories and experimental data (Morrongiello et al., 1998, Smith and Yu, 2008), which shows that infants exploit cross-situational statistics to aid the word learning process. In this way, form-referent pairs emerge by grouping the internal acoustic tokens into clusters of the same underlying unit and accumulating the associated visual features. A hierarchical agglomerative clustering (HAC) method is used for the grouping process. The HAC method initialises each element of X as separate clusters $\{C_1, \dots C_k\}$ of size 1, and then merges the two clusters C_i and C_j with the shortest distance, as defined by (4), to create $k-1$ clusters.

$$d(C_i, C_j) = \min_{v_i \in C_i, v_j \in C_j} \left[d(v_i, v_j) \right] \qquad (4)$$

This process is repeated until $d(C_i, C_j)$ is greater than the distance threshold T, leaving clusters of similar word-like segments. Table (1) displays an example of the kind of clusters that would be created by the system. The segments in bold are the cluster centroids, which is the segment with the shortest total intra-centroid distance as defined by (5).

$$\underset{v_a \in C_i}{\operatorname{argmin}} \left[\sum_j d(v_a, v_b) \right] \qquad v_b \in C_i \qquad (5)$$

LTM			
C	Segments	Visual	Accum.
1	**the_bottle_is_on_the_seat**	0 0 0 1 0 0	0 0 0 1 0 0
2	**the_bottle** the_bottle the_bottle	0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0	0 0 0 3 0 0
3	**have_you_seen_**	0 0 0 0 0 0	0 0 0 0 0 0
4	_the_b **_the_** the _	0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1 0	0 1 0 1 1 0
5	a_bath **_bath** _bath	1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0	3 0 0 0 0 0

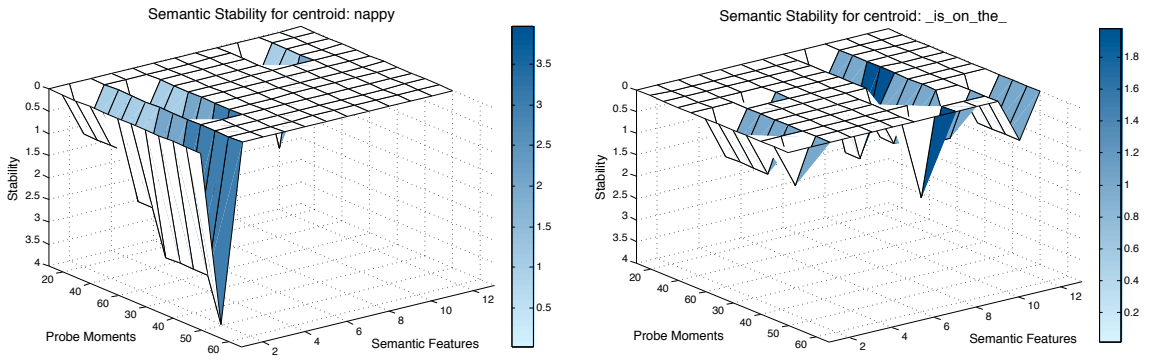
Table 1: Clusters of similar word-like units are obtained with HAC clustering. The item in bold is the *golden* representation and is semantically represented by the accumulative visual features of each token within the cluster.

With experience LA acquires a larger vocabulary of C with a greater number of representative acoustic tokens. With a greater number of exemplar acoustic tokens the system is be able to handle more variation within the speech signal. The accumulation of the visual features for each cluster also allows LA to build an increasing semantic confidence for keywords.

Table 1 shows how the word-like clusters begin to evolve. The addition of the pseudo-visual modality allows the system to derive meaning for the specific task at hand - discovering keyword units. However, it is not limited to this task as it is a general purpose pattern discovery mechanism which derives meaning through cross-situational association, which concurs with current cognitive theories of human development (Morrongiello et al., 1998, Kuhl, 2004, Smith and Yu, 2008).

2.3 Internal representations: a dynamic systems theory perspective

Describing human development as a dynamic system has become very popular within the cognitive science, where it is visualised as a continuously evolving epigenetic landscape. Current literature depicts these theoretical landscapes as hand drawn examples, such as the diagram of the evolving speech production attractor landscape displayed in figure 1.
As observed above, the landscape shows the emergence of behaviour, with varying stability, as a function of time. Each behaviour is represented as an attractor well and its stability is displayed by its depth



(a) *Epigenetic landscapes for* C_{10} *with the centroid* 'nappy'.

(b) *Epigenetic landscapes for cluster* C_{28} *with the centroid* '_is_on_the_'.

Figure 3: *These figures show the epigenetic landscapes for two of LA's internal representations. a) shows how meaning emerges with experience for clusters representing keywords, whereas b) shows that non-keywords will be semantically noisy and stay relatively flat.*

and width. As yet, there does not seem to be anyone who has attempted to visualise the emergence and evolution of the internal representations of a computational model of early language acquisition in a similar fashion.

Figure 3 displays the epigenetic landscapes for two of LA's internal representations, which are made up of a cluster of similar word-like exemplar segments. The epigenetic landscape in figure 3(a) displays a cluster with an underlying keyword representation, and the epigenetic landscape in figure 3(b) displays a cluster with a non-keyword representation. The x-axis refers to the pseudo-visual label for each of the 10 keywords, the y-axis refers to the number of utterances observed (referred to as probe moments) and the z-axis refers to the semantic stability of the cluster. Stability is simply the accumulation of each visual feature as demonstrated in table 1. Comparing the two epigenetic landscapes in figure 3 it is clear to see that clusters *not* representing a keyword are semantically noisy (fig. 3(b)). Because of this noisiness the system is not able to derive any meaning for this cluster, however, this does not mean that this cluster is not important for the language acquisition process, it just means that it hasn't been given any meaning for this particular task.

Figure 4 shows the epigenetic landscape for all internal representations in LTM, displayed as wells. The x-axis refers to the cluster space, thus, the width of each well represents the amount of acoustic variation from the median within each cluster. Each cluster is positioned in chronological order along the x-axis, with the newest being appended to the right-hand side. The y-axis refers to the probe moment, which shows the emergence and continuous evolution of each cluster after every utterance observation (only the first 12 utterances have been drawn to preserve clarity). The z-axis refers to the semantic stability (S), which is defined as the semantic cleanliness of the cluster C_i calculated using (6)

$$S = \left(\frac{\max A}{\sum A} \right) \times \max A \tag{6}$$

where A is the accumulative visual feature vector $\{a_1, \ldots, a_n\}$ for C_i.

After observing the first utterance we can see that LA stores it as an internal representation, which can then be used for recognition. It is also clear to see that the most common repetition is '`the`', as represented by the cluster with the median token '`_the_`'. It is interesting to note that although there are a lot of occurrences of this item it does not gain semantic stability. Whereas the two clusters with the median representations '`book`' and '`a_shoe`' gradually gain semantic stability, and represent keywords.

3. Experiments

3.1 Data

The training and test sets have been designed using a selection of utterances recorded within the ACORNS project. The database consists of 4000 utterances spoken by two male (M1 and M2) and two female (F1 and F2) speakers (1000 utterances per speaker). The training set consists of 450 single-speaker utterances from F1, containing both acoustic and pseudo-visual information. The test set consists of 280 single-speaker utterances from F1 that are held-out during training, and *only* contain acoustic information. The accuracy of the systems internal representations is measured with a keyword detection task, LA only observes the acoustic portion of the test utterance and must reply with the correct visual feature. Learning is incremental, therefore LA

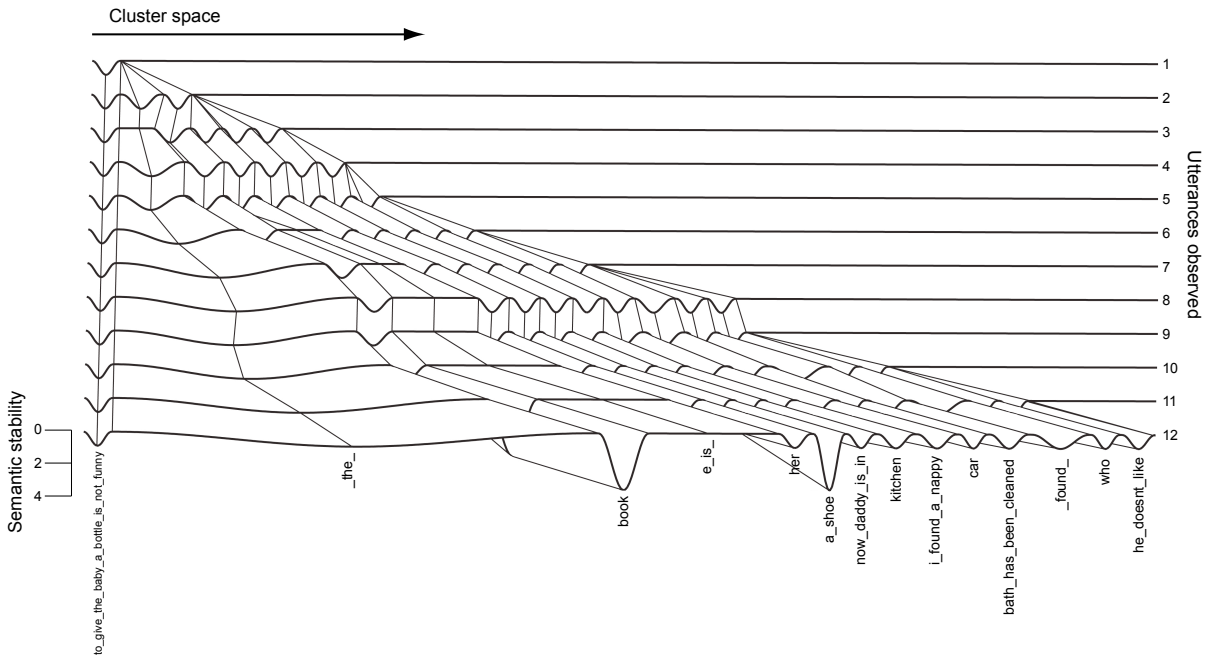


Figure 4: *Epigenetic landscape of all the internal representations during the first 12 training utterances. Each cluster is displayed as an attractor well where the acoustic variation is plotted as the width within the cluster space and semantic stability is plotted as the depth. Two clusters representing an underlying keyword have already begun to emerge from the noisy clusters -* `'book'` *and* `'a_shoe'`.

is probed after each successive training utterance is observed with the complete test set, giving us a percentage of correct keyword detections at each stage of development.

3.2 Results

Keyword detection is carried out with the acoustic DP-ngram algorithm. The test utterance is compared with each internal representation, and the visual features associated with the cluster achieving the highest quality score is replied. The system does not know a priori that each keyword is represented by only one visual feature and is penalised when replying with multiple, thus making the problem a lot more difficult but more ecologically plausible.

Figure 3.2 displays the keyword detection accuracy (y-axis) as a function of the number of utterances observed (x-axis). The green plot with circles displays the keyword detection accuracy for LA and the red dotted plot displays chance at 10%.

It can be seen from the figure that keyword representations are discovered extremely quickly but that accuracy never quite reaches 100%. This is because LA has built an internal representation of an infrequently occurring acoustic unit with an associated visual feature. This means that it will be semantically very clean and thus weighted with higher importance. A solution to this problem would be to add a forgetting mechanism in order to prune internal

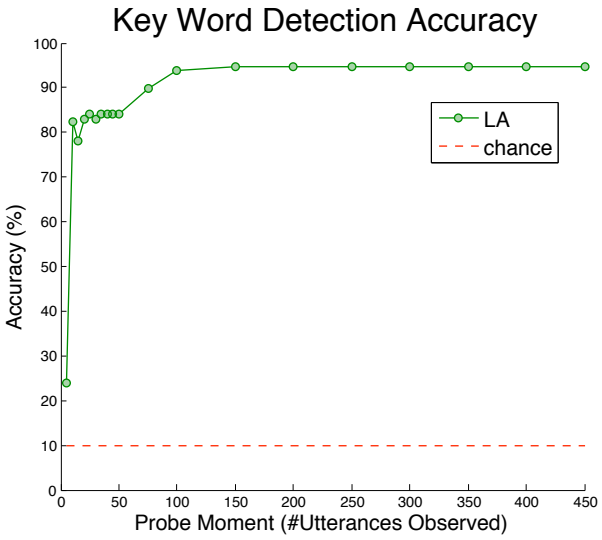


Figure 5: *Keyword detection accuracy as a function of the number of utterances observed. The green plot with circles displays LAs detection accuracy and the red dotted plot displays chance (10%).*

representations that are not useful, where usefulness is classed as how often it recognises segments.

Table 2 displays the top and bottom twelve internal representations (tables 2(a) and 2(b) respectively) out of a total of 168 that have been built within LA's LTM after observing all training utterances. The internal representations are ranked (R) in order of semantic stability (Stab), which was calculated using Eq. 6. The centroid token (Cent) of each cluster is displayed along with its cluster index (C). Observing the top twelve we can see that all of the ten keywords have emerged as the strongest clusters. It is also interesting to see the other structures that have emerged, for example multiple word units such as `'s_on_the'`, single word units such as `'_reading'` and sub-word units such as `'ing_'`.

(a) Top twelve reps

LTM			
R	C	Cent	Stab
1	27	angus	112
2	38	daddy	89
3	45	_a_bath	89
4	97	_car	84
5	34	_sho	80
6	22	_bottle	67
7	13	nappy	48
8	18	telephon	48
9	20	mummy	48
10	108	book	42
11	32	the_	23
12	7	_is_	21

(b) Bottom twelve reps

LTM			
R	C	Cent	Stab
157	11	s_on_the	0
158	24	_has_	0
159	28	s_a_b	0
160	37	_co	0
161	50	ook_	0
162	56	s_on	0
163	15	_reading	0
164	84	ing_	0
165	85	_you_	0
166	86	are_	0
167	108	_to_	0
168	111	_sits_	0

Table 2: Top (a) and bottom (b) twelve ranked clusters of 168, in order of semantic stability after observing all training utterances. It can be seen that the top 10 clusters are the keywords

4. Conclusion and discussion

In this paper, we have introduced a novel computational model of early language acquisition. Our model automatically segments speech into word-like units and derives meaning through cross-modal association. We have also presented an innovative method for comparing theoretical ideas on human development with a computational learning algorithm that is cognitively motivated. The results show that the system displays similar emergent behaviour as the DST theory of human development. Comparing the models behaviour with DST it successfully discovers keywords through self-organisation, gains knowledge without any pre-specified linguistic rules and builds internal representations which are continuously evolving with varying stability.

The results show that LA successfully builds internal representations of keywords and is able to distinguish non-keyword representations by their semantic noisiness and flat epigenetic landscape. This information would allow us to make the system more computationally efficient by reducing the size of internal representations by getting rid of or forgetting *unimportant* clusters (for this task).

Some developmental theorists believe that the DST perspective is useful for solving general problems but argue that the range of different cognitive behaviours is too great (Aslin, 1993, Port, 2000), and that it is difficult to incorporate non-observable influences such as motivation. However, for this task, the epigenetic landscape is a useful and novel tool for intuitively visualising the emergence and evolution of internal representations of a cognitively motivated computational model.

5. Further work

Experimental data shows that young language learners become faster at recognising words with experience (Swingley et al., 1999). This could be due to the development of abstract models of word representations, allowing the system to generalise. Currently, the system is using the median token of each cluster for recognition. This means that the system is building an ever increasing list of exemplar tokens, but is not taking advantage of the acoustic variation within the cluster. In order to do so it would either need to use all the tokens stored in the cluster or use a mean representation. The former is not computationally viable as the token list increases to infinity, and the latter is at the expense of accuracy. However, using a mean representation would concur with developmental data showing that infants lose the ability for finer phonetic discrimination with age.

Further work will also include the discovery of the fundamental units of speech. Theories suggest that language learners try to encode information from their environment in the most efficient way i.e. through compression (Wolff, 1982). It is hypothesised that the learner begins life discovering exemplar representations of commonly re-occurring units of speech (e.g. sentences, words, syllables etc.) and then builds prototypic models of them (i.e. an average of the units in memory). LA attempts to learn in the most efficient way, therefore, patterns are discovered from a large to small scale. This means that during the early stages of language development, the infant will predominantly use internal representations of sentences and words before it has an optimised lexicon for its native language. We believe that the word-spurt phenomena would be replicated in our model with this learning mechanism in place. When the system has a robust lexicon of the fundamental units then new words can be composed by concatenating these models rather than starting with an ever-increasing list of exemplar units.

Acknowledgements

This research was funded by the European Commission, under contract number FP6-034362, in the ACORNS project (www.acorns-project.org).

References

Aimetti, G. (2009). Modelling early language acquisition skills: Towards a general statistical learning mechanism. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 1–9. Association for Computational Linguistics.

Aslin, R. N. (1993). *Dynamic Systems in Development: Applications*, chapter The strange attractiveness of dynamic systems to development, pages 385–399. Cambridge, MIT Press.

Chomsky, N. (1975). *Reflections on Language*. New York: Pantheon Books.

Evans, J. L. (2007). *Blackwell Handbook of Language Development*, chapter 7. The Emergence of Language: A Dynamical Systems Account, pages 128–147. Blackwell Publishing.

Jones, D. M., Hughes, R. W., and Macken, W. J. (2006). Perceptual organization masquerading as phonological storage: Further support for a perceptual-gestural view of short-term memory. *Journal of Memory and Language*, 54(2):265–281.

Kelso, J. A. S. (1995). The self-organization of brain and behavior. In *Dynamic Patterns*, chapter 2, pages 46–53. MIT Press.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature*, 5:831–843.

Morrongiello, B. A., Fenwick, K. D., and Chance, G. (1998). Crossmodal learning in newborn infants: Inferences about properties of auditory-visual events. *Infant Behavior and Development*, 21(Index to Volume 21, Issue 4):543 – 553.

Muchisky, M., Gerschkoff-Stowe, L., Cole, E., and Thelen, E. (1996). The epigenetic landscape revisited: A dynamic interpretation. *Advances in Infancy Research*, 10:121–159.

Newell, M. K., Liu, Y.-T., and Mayer-Kress, G. (2003). A dynamical systems interpretation of epigenetic landscapes for infant motor development. *Infant Behavior and Development*, 26(4):449–472.

Nowell, P. and Moore, R. K. (1995). The application of dynamic programming techniques to mon-word based topic spotting. *EuroSpeech '95*, pages 1355–1358.

Park, A. and Glass, J. (2008). Unsupervised pattern discovery in speech. In *Trans. ALSP*, volume 16, pages 186–197.

Pinker, S. (1994). *The Language Instinct*. New York: Morrow.

Port, R. F. (2000). Dynamical systems hypothesis in cognitive science. In *Encyclopedia of Cognitive Science*.

Räsänen, O. J., Laine, U. K., and Altosaar, T. (2009). A noise robust method for pattern discovery in quantized time-series: the concept matrix approach. In *Interspeech 2009*.

Sankoff, D. and Kruskal, J. B. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, chapter Finding similar portions of two sequences, pages 293–296. Addison-Wesley Publishing Company, Inc.

Savelsbergh, G. J. P. and Van der Kamp, J. (1993). The development of coordination in infancy. *Advances in Psychology*, 97:289–317.

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

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

Stouten, V., Demuynck, K., and Van hamme, H. (2007). Automatically learning the units of speech by non-negative matrix factorisation. In *Interspeech 2007*.

Swingley, D., Pinto, J. P., and Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition*, 71:73–108.

ten Bosch, L. and Cranen, B. (2007). A computational model for unsupervised word discovery. In *INTERSPEECH 2007*.

Thelen, E. and Fisher, D. M. (1982). Newborn stepping: An explanation for a "disapearing rflex". *Developmental Psychology*, 18:760–775.

Thelen, E. and Smith, L. B. (1995). A dynamic systems approach to development of cognition and action. *Journal of Cognitive Neuroscience*, 7(4):512–514.

Waddington, C. H. (1957). *The strategy of the genes*. London: Allen & Unwin.

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language and Communication*, (2):57–89.